

2016

Confluence of Vision and Natural Language Processing for Cross-media Semantic Relations Extraction

Amara Tariq
University of Central Florida



Part of the [Computer Sciences Commons](#)

Find similar works at: <https://stars.library.ucf.edu/etd>

University of Central Florida Libraries <http://library.ucf.edu>

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2004-2019 by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation

Tariq, Amara, "Confluence of Vision and Natural Language Processing for Cross-media Semantic Relations Extraction" (2016). *Electronic Theses and Dissertations, 2004-2019*. 5239.
<https://stars.library.ucf.edu/etd/5239>



CONFLUENCE OF VISION AND NATURAL LANGUAGE PROCESSING FOR
CROSS-MEDIA SEMANTIC RELATIONS EXTRACTION

by

AMARA TARIQ
M.S. University of Central Florida, 2014

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2016

Major Professor: Hassan Foroosh

© 2016 Amara Tariq

ABSTRACT

In this dissertation, we focus on extracting and understanding semantically meaningful relationships between data items of various modalities; especially relations between images and natural language. We explore the ideas and techniques to integrate such cross-media semantic relations for machine understanding of large heterogeneous datasets, made available through the expansion of the World Wide Web. The datasets collected from social media websites, news media outlets and blogging platforms usually contain multiple modalities of data. Intelligent systems are needed to automatically make sense out of these datasets and present them in such a way that humans can find the relevant pieces of information or get a summary of the available material. Such systems have to process multiple modalities of data such as images, text, linguistic features, and structured data in reference to each other. For example, image and video search and retrieval engines are required to understand the relations between visual and textual data so that they can provide relevant answers in the form of images and videos to the users' queries presented in the form of text.

We emphasize the automatic extraction of semantic topics or concepts from the data available in any form such as images, free-flowing text or metadata. These semantic concepts/topics become the basis of semantic relations across heterogeneous data types, e.g., visual and textual data. A classic problem involving image-text relations is the automatic generation of textual descriptions of images. This problem is the main focus of our work. In many cases, large amount of text is associated with images. Deep exploration of linguistic features of such text is required to fully utilize the semantic information encoded in it. A news dataset involving images and news articles is an example of this scenario. We devise frameworks for automatic news image description generation based on the semantic relations of images, as well as semantic understanding of linguistic features of the news articles.

to my brother Umer Tariq

Thank you for all your support and guidance

ACKNOWLEDGMENTS

I would like to acknowledge my appreciation for my doctoral adviser, Dr. Hassan Foroosh. Without his guidance and encouragement, i would not have been able to explore and develop my research interests.

I would also like to express my appreciation for Dr. Asim Karim who supervised my Masters' thesis and also contributed to my doctoral research. His guidance was necessary to get me started on research in machine learning and knowledge mining.

My sincere appreciation is extended to the members of my dissertation committee, Dr. Marianna Pensky, Dr. GuoJun Qi, and Dr. Avelino Gonzalez. Their contribution was crucial to timely completion of this dissertation.

My deepest gratitude is reserved for my parents for their never-ending love and support. Also, special thanks to my younger brother Kamran for keeping the humor alive though all the ups and downs of life.

TABLE OF CONTENTS

LIST OF FIGURES	xii
LIST OF TABLES	xvii
CHAPTER 1: INTRODUCTION	1
1.1 Automatic Image Annotation	2
1.1.1 News Image Annotation	6
1.1.2 News Image Caption Generation	9
1.2 Semantic Relations between Named Entities	11
CHAPTER 2: RELATED WORK	14
2.1 Image Annotation and Caption Generation	14
2.1.1 News Image Annotation and Caption generation	20
2.2 Semantic Network of Named Entities	22
CHAPTER 3: AUTOMATIC IMAGE ANNOTATION	27
3.1 Scene-based Automatic Image Annotation	28
3.1.1 System Architecture	29

3.1.1.1	Scene Categorization	30
3.1.1.2	Relevance Model based Image Annotation	32
3.1.2	Evaluation	34
3.1.2.1	Datasets	34
3.1.2.2	Visual Features	35
3.1.2.3	Results	35
3.1.2.4	Cluster Expansion for Large Datasets	37
3.2	Feature-Independent Semantic Relations Extraction for Image Annotation	39
3.2.1	System Architecture	39
3.2.1.1	Semantic Relation Extraction Framework	40
3.2.1.1.1	Semantic Concept-based Categories	40
3.2.1.1.2	<i>Semantic Signature</i> through Tensor Analysis	42
3.2.1.1.3	Semantic Relations through Tensor Analysis	43
3.2.1.2	Relevance Model based Image Annotation	45
3.2.2	Evaluation	47
3.2.2.1	Results	48
3.2.2.2	Implications of Tensor Decomposition	50

3.2.2.3	Computational Complexity	52
3.3	Multi-Layer Sparse Coding Framework for Image Annotation	53
3.3.1	System Architecture	56
3.3.1.1	Visual Feature Extraction	57
3.3.1.2	Theme-based Clustering	58
3.3.1.3	Multi-Layer Sparse Coding	58
3.3.1.3.1	Group sparse coding for theme identification	59
3.3.1.3.2	Regularized linear regression modeling for tag prediction	60
3.3.2	Evaluation	62
3.3.2.1	Results	63
3.3.2.2	Noise Reduction	67
3.3.2.3	Time Complexity	68
3.3.2.4	Theme Selection and Image Organization	69
3.3.2.5	Precision for Descriptive Words	70
3.4	Conclusion	72
CHAPTER 4: CROSS-MEDIA SEMANTIC RELATIONS FOR NEWS MATERIAL . . .		74
4.1	News Dataset	76

4.2	News Image Annotation	79
4.2.1	Context-sensitive News Image Annotation System	80
4.2.1.1	Context Estimation	81
4.2.1.1.1	Scene Characteristics of Images	82
4.2.1.1.2	News Articles	84
4.2.1.1.3	News Category Labels	86
4.2.1.1.4	Article Keywords	88
4.2.1.1.5	Combination of Heterogeneous Context Sources	90
4.2.1.2	Context-sensitive Generative Model	91
4.2.2	Evaluation of News Image Annotation System	93
4.2.2.1	Comparison Models	93
4.2.2.2	Results	94
4.2.2.3	Observations	94
4.2.2.4	Parameter optimization	98
4.2.2.4.1	Manual Tuning	99
4.2.2.4.2	Least Squares Error Minimization	99
4.3	News Image Caption Generation	103

4.3.1	Context-sensitive News Image Caption Generation System	104
4.3.1.1	Context-sensitive Word Distribution from Image	105
4.3.1.2	Context-sensitive Word Distribution from Article	105
4.3.1.3	Extraction of Caption	106
4.3.2	Evaluation of News Image Caption Generation System	107
4.3.2.1	Comparison Models	109
4.3.2.2	Results	111
4.3.2.3	Observations	111
4.3.3	Study of <i>Semantic Gap</i> in News Images	116
4.4	Semantic Network of Named Entities	119
4.4.1	Sparse Structured Modeling for Named Entities Relations Extraction . . .	122
4.4.1.1	Semantic topics	123
4.4.1.1.1	Co-occurrence-based word groups	123
4.4.1.1.2	Keyword-based word groups	124
4.4.1.1.3	Topic-based word groups	126
4.4.1.2	Sparse Structured Modeling	126
4.4.2	System Output	130

4.4.2.1	Network of Named Entities	130
4.4.2.2	News Events	132
4.4.2.3	Dynamics of Relations	132
4.4.3	Evaluation for Named Entity Relations	135
4.4.3.1	Wikipedia-based Evaluation	136
4.4.3.2	Retrieval-based Evaluation	137
4.4.3.3	Automatic Evaluation Results	139
4.4.3.4	Effects of Word Group Formation Methods	143
4.4.3.5	Human Evaluation Study	144
4.4.3.6	Co-occurrence-based Baseline Model	148
4.4.3.7	Value of Sparse Group Learning	148
4.4.3.8	Sensitivity Analysis	150
4.4.3.9	Time Complexity and Scalability	152
4.5	Conclusion	152
CHAPTER 5: CONCLUSION		155
LIST OF REFERENCES		161

LIST OF FIGURES

3.1	A sample <i>Scene</i> from IAPR TC-12 dataset.	31
3.2	A sample <i>Scene</i> from IAPR TC-12 dataset.	31
3.3	An example of semantic theme formed on the basis of similarity in textual descriptions	40
3.4	An example of semantic theme formed on the basis of similarity in textual descriptions	40
3.5	Tensor formation: images of one group are stacked together to form one tensor	42
3.6	Comparison of rank-1 Tucker decomposition with visually similar and dissimilar image inserted into a tensor; Blue curve: Original Tucker decomposition vector R , Green curve: New Tucker decomposition vector R^Y with image Y inserted into the <i>semantic tensor</i> \mathbb{T}_c such that Y is visually similar to the images already contained in \mathbb{T}_c , Red curve: New decomposition vector R^Y with image Y inserted into \mathbb{T}_c such that Y is visually dissimilar to the images of \mathbb{T}_c	44
3.7	Tucker decomposition: $\mathbf{U} = words \times word\text{-}groups$, $\mathbf{V} = authors \times author\text{-}groups$, $\mathbf{W} = keywords \times keyword\text{-}groups$, $R1$, $R2$ and $R3$ represent word, author and keyword groups	50

3.8	Rank-1 Tucker decomposition: S is a scalar, P , Q and R are vectors, $R = Image-indices \times 1$ where 1 represent the single <i>context</i> group represented by tensor.	51
3.9	System Architecture; \mathcal{X} denotes the set of training items. \mathcal{X}^y represents the subset of training items that belong to <i>themes</i> selected for test image Y . V_Y is the set of words selected for test image Y	56
3.10	A sample <i>theme</i> from Flickr30K dataset; ‘ person playing banjo’ seems to the distinctive characteristic of this <i>theme</i>	58
3.11	Sample images from multiple <i>themes</i> associated with the given test image . . .	70
3.12	Precision vs. Frequency of MBRM and MultiSC-AIA (IAPR TC-12)	71
4.1	MSCOCO image-caption pairs	80
4.2	IAPR TC-12 image-caption pairs	80
4.3	TIME image-caption pairs	80
4.4	Visually similar images - Different news stories.	84
4.5	Similar articles - Visually different images	84
4.6	Distribution of items among news categories for TIME dataset	88
4.7	Distribution of items among article keywords for TIME dataset	89
4.8	TIME image-caption pairs; CNN-assigned ImageNet labels are written in bold face.	116

4.9	News image from the TIME dataset; Ground-truth caption is “ <i>Facebook CEO Mark Zuckerberg</i> , NeuralTalk-generated caption is “ <i>A man is sitting on the rock</i> , Caption generated by <i>context-EXT</i> is “ <i>Facebook was originally not created to be a company, Zuckerberg wrote in Facebooks prospectus.</i>	118
4.10	Distributions of named entities over time; each bar represents frequency of a named entity during one month.	121
4.11	Semantic network of named entities for Nov-Dec 2012 (TIME dataset)	131
4.12	Semantic network of named entities for Jun-Jul 2013 (TIME dataset)	131
4.13	Variation of average <i>strength</i> and WLM of relations over time (TIME dataset). Relations among the following named entities exist in each time period: {Syria, Bashar Asad, Cairo, Damascus, Jerusalem, Hamas, Gaza, Israel Egypt, Benghazi, Hillary Clinton}; x-axis: Time period (months). y-axis: Mean WLM and relation <i>strength</i>	133
4.14	Effect of threshold γ on evaluation measures on TIME dataset; x-axis: Time period (month), y-axis: Evaluation measure (connectivity, mean WLM, mean retrieval score, mean <i>strength</i>); Solid line: $\gamma = \gamma_1$, Dotted line: $\gamma = \gamma_2$ where $\gamma_1 < \gamma_2$; Mean of each curve given in legends.	140
4.15	Summary statistics (boxplots) for connectivity, mean WLM, mean retrieval score, and mean <i>strength</i> at γ_2 minus that at γ_1 over all time periods (TIME dataset), $\gamma_2 > \gamma_1$	141

4.16	Effect of threshold γ on evaluation measures on BBC dataset; x-axis: Dataset sample, y-axis: Evaluation measure (connectivity, mean WLM, mean retrieval score, or mean <i>strength</i>); Solid line: $\gamma = \gamma_1$, Dotted line: $\gamma = \gamma_2$ where $\gamma_1 < \gamma_2$; Mean of each curve given in legends.	142
4.17	Summary statistics (boxplots) for connectivity, mean WLM, mean retrieval score, and mean <i>strength</i> at γ_2 minus that at γ_1 over all samples (BBC dataset), where $\gamma_2 > \gamma_1$	143
4.18	Human evaluation of NELasso; height of each bar represents the mean of human-assigned strength to the relations discovered by NELasso; Blue: $\gamma = \gamma_1$, Red: $\gamma = \gamma_2$ such that $\gamma_1 < \gamma_2$	145
4.19	Fraction of human-identified relations discovered by NELasso; x-axis: Minimum $strH(r_{ij})$ of human-identified relations, y-axis: fraction of human-identified relations discovered by NELasso; Blue: $\gamma = \gamma_1$, Red: $\gamma = \gamma_2$ such that $\gamma_1 < \gamma_2$	146
4.20	Comparison between co-occurrence-based baseline model and various configurations of NELasso using WLM (TIME dataset); x-axis: Time period (month), y-axis: mean WLM; Mean of each curve given in legends.	147
4.21	Effect of threshold ζ of linear model baseline system on evaluation measures (TIME dataset); x-axis: Time period (month), y-axis: Evaluation measure (connectivity, mean WLM, mean retrieval score, mean <i>strength</i>); Solid line: $\zeta = \zeta_1$, Dotted line: $\zeta = \zeta_2$ where $\zeta_1 < \zeta_2$; Mean of each curve given in legends.	147

4.22	Comparison between NELasso and linear model baseline system (TIME dataset); x-axis: Time period (month), y-axis: Evaluation measure (connectivity, mean WLM); Mean of each curve given in legends.	149
4.23	Comparison between co-occurrence and linear model based baselines on the basis of WLM (TIME dataset); x-axis: Time period (month), y-axis: mean WLM for relations; Mean of each curve in legend.	149
4.24	Effects of threshold γ on the consistency of system's output (TIME dataset) .	151

LIST OF TABLES

3.1	Performance evaluation for IAPR TC-12 dataset	36
3.2	Performance evaluation for ESP game dataset	36
3.3	Performance evaluation for ESP-large dataset	38
3.4	Performance evaluation for IAPR TC-12 dataset	49
3.5	Performance evaluation for ESP-game dataset	49
3.6	Sample of words with low and high recall values	49
3.7	Test images and training images from multiple <i>themes</i> related to them; sam- ples from different <i>themes</i> are separated by bold vertical lines.	61
3.8	Performance evaluation for IAPR-TC-12 dataset	64
3.9	Performance evaluation for ESP game dataset	65
3.10	Performance evaluation for Flickr30K dataset	66
3.11	Sample high and low recall words from three all datasets; The words with high frequency (e.g., ‘man’, ‘woman’, ‘dog’) or related to distinctive visual <i>themes</i> (e.g., <i>tennis match</i> or <i>bicycle race</i>) achieve better recall.	66
3.12	Noise in training data for individual words	68
4.1	News categories of a few popular news sources	87

4.2	Evaluation of <i>context</i> -AIA and comparative annotation models (BBC dataset)	95
4.3	Performance evaluation of previously proposed annotation methods (TIME dataset)	96
4.4	Baseline annotation performance (TIME dataset)	96
4.5	Comparative performance of <i>context</i> sources of <i>context</i> -AIA (TIME dataset)	96
4.6	Performance of combinations of <i>context</i> sources of <i>context</i> -AIA (TIME dataset)	97
4.7	Comparative quality of <i>context</i> source	103
4.8	Performance of caption generation systems over BBC dataset (Average length of ground truth caption is 10); The best score in each block is indicated by bold font.	112
4.9	Performance of caption generation systems over TIME dataset. Average length of ground truth caption is 20; The best score in each column is indicated by bold font. [†] :Significantly different from <i>context</i> -EXT in terms of METEOR. *:Significantly different from <i>phrase</i> -ABS(<i>context</i> -AIA) in terms of METEOR.	112
4.10	Characteristics of <i>context</i> -EXT framework; The underlined words in system-generated captions overlap with corresponding ground truth captions.	113
4.11	Comparative analysis of captions generated by various systems	114
4.12	Visual features comparison for image annotation (‘conv5’: last convolutional layer of CNN, ‘fc7’: last fully connected layer of CNN, ‘grid’: grid-based visual features)	117

4.13	Example cliques discovered by our system (TIME dataset); each clique corresponds to a distinct news event indicated by the <i>type</i> of the relation	132
4.14	Example of a named entity involved in different relations over time (TIME dataset). The <i>strength</i> of the relation using keyword based word groups is shown in parenthesis	134
4.15	Sample relations with more than one relation <i>type</i> in one time period (TIME dataset)	134

CHAPTER 1: INTRODUCTION

With the expansion of the World Wide Web, people are gaining access to larger and more diverse datasets than ever before. Social media websites, online pages of news media outlets, and blogging platforms are the hosts of huge amount of information in the form of large and diverse datasets. Data collected from such hosts contains various data modalities such as images, short or long sequences of text, videos, audio, keywords, timestamps, etc. It is hard for humans to quickly extract relevant pieces of information from such data sources. Hence, automatic or machine understanding of such datasets is essential to fully explore and utilize their information-rich contents. There are vast potential applications for systems that can quickly extract relevant pieces of information from such collections of data, or summarize and visualize their contents in a way that are easier for humans to understand.

Effective tools for automatically understanding large and diverse datasets need to process various modalities of data in reference to each other to extract meaningful information. There are various examples of tools and applications that require simultaneous processing of multiple modalities of data. Most visual search engines are provided with users' queries in the form of text and they have to retrieve visual information, i.e., images or videos, relevant to the query. News websites are a source of information-rich data in the forms of textual news articles, news images, timestamps of news or events, news categorization keywords, etc. Automatic systems for tracking, summarizing and linking similar news stories together can be extremely beneficial for readers as well as journalists researching news events. Such systems need to understand semantic relations between various data types involved in news collections, i.e., images, text, metadata, etc.

In this dissertation, we explore ideas and techniques to automatically develop semantically meaningful understanding of multi-modality datasets with a focus on image-text relations. The core

of our approach is the automatic extraction of meaningful semantic topics and concepts from the available data. We devise various frameworks to group together items of certain data type in such a way that each group is a representative of one semantic topic or concept. For example, a group of words may define a ‘topic’ or a group of images may represent a semantically meaningful ‘scene-category’. Data items of different modality such as images and text are linked to each other through their relations to common semantic topics. Hence, each cross-modality link has semantic significance and plays an important role in automatic understanding of the dataset. Semantic relations between data items are the key to expressing semantic information regarding multi-modality datasets in a way that is useful for information search, retrieval, organization, summarization and visualization tools.

In the following sections, we introduce our ideas for the solution of a classic problem involving image-text relations, i.e., automatic generation of textual descriptions of images. Such descriptions can be in the form of sets of independent words or sentences. In general, image-text datasets involve only short sequences of text associated with images. Though, news websites have become a popular source of information-rich datasets which contain long sequences of natural language in the form of news articles associated with images. Deep exploration of linguistic features is required to fully comprehend the semantic information encoded in the text of news article. We also introduce the problem of machine understanding of such text and our approach to develop automatic semantic understanding of important linguistic features.

1.1 Automatic Image Annotation

Search and retrieval engines need to understand the relations between textual queries provided by users and various modalities of data. Image search engines generally rely on textual tags associated with images to find images relevant to the textual input query. It is expected that people will

provide such tags with images when they upload them. This is not always the case. People may provide incomplete tags or may not provide tags at all when they upload images. The search engine sifts through the text of the web page the image is available on, in the absence of such tags, to establish potential links between the query and the image. Sifting through the text on the webpage generates noisy textual tags for these images. An automatic system to efficiently tag images with concise and accurate textual descriptions in the form of individual words or sentences can be extremely beneficial for image search and retrieval engines. Such systems are called *automatic image annotation* systems and development of such systems has been widely studied by image processing, computer vision, as well as natural language processing research communities.

The main challenge for any automatic image annotation system is to build meaningful relations between two different modalities of data, i.e., images and text. Traditionally, images are represented by their low-level visual features such as mean and standard deviation of their color channels, histogram of gradients or the energy of edge-detection filters such as Gabor filter. Text is generally represented in the form of frequency of words or binary values representing presence or absence of any word. Such representation schemes do not have any strict one-to-one correspondence between them. The lack of correlation between textual and visual features is termed as the *semantic gap*. More complex visual features, such as *blobs* (portions of images with uniform color and texture properties) or SIFT features (local image features detected and described through scale-invariant feature transform (SIFT) algorithm[72]), have been explored to bridge this gap. *Bag-of-words* (BoW) scheme is utilized to employ such complex features. Continuous feature vectors of *blobs* or SIFT are clustered to form a limited number of clusters, called *visual words*. Each *blob* or SIFT feature vector of each image is mapped to one of these clusters/*visual words*. Image is represented by a vector of discrete values where each entry indicates presence/absence or frequency of one *visual word*. Hence, these complex visual features are transformed from continuous to discrete domain. This process introduces quantization error. Objects and actions depicted in images have

also been used as visual features. Such image representations rely heavily on the detection and the recognition systems developed for various applications of computer vision. These detection and recognition systems do not only add to the time and computational complexity of the over all system, but are also limited in availability for unconstrained and practical settings. Recently, deep convolutional neural networks (CNN) have gained tremendous popularity for the task of object detection. CNN frameworks are effective as they produce representations for raw images without needing manual crafting of visual features. CNN requires large labeled datasets for training. If the image annotation system employs a pre-trained CNN framework, the performance of the system heavily depends on how similar (or dissimilar) in nature test images are to the images in the database used for pre-training the CNN. In other words, annotation system may suffer the challenges studied under the field of *transfer learning*. *Transfer learning* is the study of the problems arising when systems are tested over datasets which are of different nature from the datasets used to train them[91].

A variety of modeling schemes have been employed to predict suitable words for images represented in terms of various types of visual features. Relevance model from the domain of machine translation has been adapted to the task of *translating* visual features into words. This type of modeling executes an expectation process over the training data to estimate joint probability of the words and the visual features. Nearest-neighbor type algorithms have also been employed to predict words for images. The core idea of this approach is to find the most similar or the ‘nearest’ neighbors of any test image in the set of training images. Later, words associated with these ‘nearest’ neighbors in the training set are transferred to the new image. Names of objects and actions represented in images are also used as textual tags of the image. Some systems form sentences using names of objects, actions and image characteristics as nouns, verbs and adjectives, respectively. Section 2.1 in Chapter 2 describes various image annotation approaches in detail. Chapter 3 presents several image annotation systems that we devised based on the core philosophy of this

dissertation, i.e., automatic understanding of semantic relation across different data modalities.

Image annotation systems are generally provided with training datasets consisting of image-description pairs. We tackle the problem of *semantic gap* by extracting semantically meaningful concepts and establishing relations between training images and these concept. The main idea is that the individual low-level visual feature may not correspond to any word directly, but meaningful classes of images can be formed such that all images in one class are representatives of the same semantic concept. When a new image is presented whose textual description needs to be generated, the first step is to establish its association with image classes present in the training images. It is easier to establish similarity between the test image and the training image classes as these items belong to the same modality of data, as compared to directly establishing similarity between the test image and words. The new image is, in turn, associated with the semantic concept that is linked to the image class it belongs to. This association is further exploited as prior knowledge while predicting words to describe the new image. In this dissertation, we explain various ways that we employed to extract semantic concepts from the set of image-description pairs, and from even more heterogeneous datasets involving images, long and short text sequence and structured data labels.

We employed scene-representations of training images to cluster images depicting similar scenes. Studies have shown that humans can identify scenes presented in the image without recognizing individual objects shown in images[94, 6]. Oliva et. al. proposed a model that recognizes scenes while bypassing image segmentation and individual object detection[86]. Recently, deep convolutional neural networks (CNN) have been trained over a large database of images, called ‘Places’[129]. Each image in Places database is labeled by the scene it represents. Image representations learned from such CNN are very effective for the prediction of scene-type labels contained in the Places database. Since image annotation is aimed at describing the details of image contents, it can benefit from the information about the type of scene the image represents. For example, images depicting scenes of busy streets of big cities are more likely to show objects like ‘cars’, ‘tall

buildings’, ‘people’, etc. The images that show scenes of countryside are more likely to show stretches of ‘grass ’or clear blue ‘sky’. Therefore, we employed scene representations of images (GIST feature vectors and image representations learned from Places-trained CNN) to form scene categories (Figures 3.1 and 3.2 in Section 3.1 of Chapter 3). The relations between news images and these scene categories is used as prior knowledge in the process of image annotation.

A semantic concept or a topic can be represented in terms of words that express or are related to that concept/topic. Image annotation systems are generally provided with images whose textual labels are known, for training purposes. The training images associated with words belonging to the same concept can be treated as the visual representatives of that concept. We devised a method to cluster training images in such a way that the images in one group share the same *distinctive* and *descriptive* words in their descriptions (Figures 3.3 and 3.4 in Section 3.2 of Chapter 3). The system should focus on shared words which not only correlate with the visual contents, but also set the image group apart from other images. The words that are associated with too many images do not provide any *distinctive* information. We turned to the *tfIdf* (term frequency-inverse document frequency) representation of text from the field of text mining. Such representation allows the system to focus on words with high information content, rather than the extremely frequent words of low information content. We devised multiple methodologies to establish relations between a test image and semantic concepts represented by *distinctive* words. Such relations are later employed as prior knowledge while annotating images with individual words.

1.1.1 News Image Annotation

News datasets collected from the websites of news media outlets are excellent examples of collections in which multiple data types are available and there is correspondence between different data types. A news article and its associated image, keyword, category labels and timestamp have

some connection to each other. Our work is focused extracting semantically meaningful relations between different data types, e.g.g, images and text, and employing such relations for machine understanding of large heterogeneous datasets. Therefore, we focus on news datasets and start by devising an automatic description generation system for news images.

The difference between standard image annotation and news image annotation is that the news images are, in a sense, already annotated with a variety of information. Each image can be considered as ‘annotated’ with the article it accompanies, as well as the title, keywords or timestamps of that article. The goal of the annotation system is to automatically predict words that describe an image. In the case of news images, the predicted words should match the words used in real world image captions written by news editors. The nature of such captions is different from the nature of news articles. A news article may discuss many stories or many aspects of the same story. A caption is generally a concise description of the image in reference to the related story mentioned in the corresponding news article. We call the problem of predicting individual words for description of news images as ‘annotation’ to comply with the terminology of previously published papers.

Automatically predicting words that match the ground truth caption is challenging, but such an annotation system can replace human caption writers if these words are transformed into sentences, and make the job of news article writers easier. News search, retrieval, organization and summarization systems can benefit from such an annotation system. A news image annotation system will also have to build essential relations between news images and text that can be invaluable to any news tracking, summarization and retrieval system.

News image captions are different from the descriptions of images in standard image annotation benchmark datasets. Image descriptions in standard image annotation datasets strictly describe the visual contents of images without providing any reference or *context*. It is so because such image descriptions have been carefully crafted by human annotators to describe image contents for

evaluation of image annotation systems. real world image descriptions have not been commonly used for evaluation. In reality, people uploading photos of vacation or celebratory events on social media websites, describe such images in reference to their occasion. Popular image annotation datasets like Flickr30K[127] and MSCOCO[70] are collections of such social media images. real world captions for images in these collections were ignored. Human annotators with no knowledge regarding the background of these images, were asked to write captions to describe the visual contents of images. These ‘artificial’ captions which are completely devoid of *context* of images, are used as ground truth. On the other hand, news images are associated with their real world descriptions or captions. Captions describe image contents in reference to the story behind the image, usually presented in the accompanying articles. The task of news image annotation is to automatically produce image annotations to match these real world captions. Figures 4.1, 4.2 and 4.3 show sample image-description pairs from the standard image annotation datasets, as well as news image dataset. These figures highlight the difference in nature of the real world captions and that of artificial image descriptions.

Past research in natural language processing community has suggested that given the vast amount of text available with news images, there is no need to consider the visual information to produce descriptions for such images. If standard image annotation schemes are applied to news images annotation, additional information hidden in accompanying news articles and related metadata is essentially ignored. Figures 4.4 and 4.5 show pairs of news images and their ground truth captions. These figures show that both image contents and the contents of articles have significant effects over appropriate image captions. Similar-looking images can have different captions because of the different events discussed in their respective articles. Even when two articles discuss the same topics, their associated images can have different captions because of the difference in their visual contents. A few previously published papers have tried to combine information from both articles and images to produce image captions. The main challenge is finding a common representation

scheme for the two distinct information sources. Feng et al. proposed the use of *visual words* or BoW representation scheme for images while textual contents of articles are already in the form of *words*[31, 32]. As explained earlier, such discrete representation scheme for images introduces quantization error in the annotation system.

We argue that the inclusion of the *context* of the image in the ground truth image description results in widening of the *semantic gap* between visual and textual features. Not only there is no strong correspondence between visual and textual features, but also visual similarity does not guarantee similarity in textual features. News images also have associated information sources like articles, keywords, etc., which are not present for standard image annotation datasets. We devise frameworks to utilize information not only from articles and images, but from every resource available. We adapt our core idea of establishing relations between images and semantic concepts to include a diverse set of semantic information sources, i.e., news article, news category labels, article keywords, and scene information of images. These information sources belong to different data types. Propagation of semantic information between such heterogeneous information sources requires the use of some common ‘representation space’ for the semantic information. We employ *probability space* as the common representation space. Semantic relations estimated from each source can be transformed into a *probability distribution* across all available semantic concepts for each image. Our system employs the aggregated semantic information from all sources as prior knowledge when individual words for images are to be predicted.

1.1.2 News Image Caption Generation

The next task is to transform individual words selected for each image into a properly formed sentence. This task is named as *caption generation*. Various approaches have been used in the past for generating captions from words for both standard image annotation datasets as well as

news images. Templates of sentences may be filled with appropriate nouns, verbs and adjectives selected through image annotation process. Language modeling has been employed to generate sentences conditioned over the association between images and words. Recurrent neural networks (RNN) and long short-term memory networks (LSTM) have been employed recently to produce sentence-like captions.

The systems producing captions for news images are at an advantage as large amount of grammatically-correct text is available with images in the form of articles. It is highly likely that at least some portion of this text correlates directly with the associated image. Therefore, the problem of news image caption generation can be modeled as an *extraction* process that selects the most suitable sentence from the article as the caption of the image. Appropriate words selected for the image by the annotation system are critical to this extraction process. This kind of process is similar to the *extractive* summarization approach used for textual documents. As opposed to the *abstractive* summarization approach that ‘generates’ sentences to summarize the document, *extractive* approach simply ‘extracts’ the best set of sentences that briefly cover all aspects of the textual contents of the document[52, 78]. *Extractive* approach based systems are more efficient in terms of time and computational complexity as compared to the systems based on *abstractive* approach.

We use an *extractive* approach for generation of news image captions. Our framework estimates a probability distribution over words for the image caption. This distribution is the combination of two probability distributions; 1) the distribution over words estimated by the image annotation system, and 2) the probability of words being present in the caption conditioned over vocabulary of the accompanying article. Semantic relational information is inherently part of the first word distribution as the image annotation system relies heavily on this information. We also devise a method to estimate the second word distribution in a manner that incorporates semantic relational information. The sentence in the article whose word distribution matches the most closely to the estimated word distribution for caption, is extracted to serve as image caption.

1.2 Semantic Relations between Named Entities

The availability of long sequences of natural language in the form of news articles with news images opens up various possibilities for mining meaningful semantic information from careful examination of news articles. Such articles contain words from a very large vocabulary set but some of these words form linguistic features with special meaning. Named entities are an example of such linguistic features. Named entities are the words that indicate the names of people, places and organizations, mentioned in the free-flowing text. Named entities constitute a very important part of news articles and blogs. A study has shown that named entities are the most commonly used words as search queries for blog search engines[82]. It is understandable as news articles and blogs usually discuss main political, sports, business or entertainment-related events which are taking place somewhere (named entity type: place), involve some people (named entity type: person) or institutions (named entity type: organization). The time-line of these events is also very important but this information is usually incorporated in the timestamps of news articles or blogs.

Image-text relations can be semantically enriched if correspondence between images and such special linguistic features can be made. Various attempts have been made in the past to establish links between visual data and named entities. The problem of identification of faces of people (person)[5, 89, 110, 46] and recognition of landmarks (places)[16, 17, 1, 44] and logos (companies)[100, 14, 29, 55, 101] in images has been widely studied in computer vision and image processing communities. There is only limited amount of semantic information readily available from images, resulting in shallow understanding of such entities. On the other hand, free-flowing text like news articles encodes vast semantic information about these entities. We devise a framework for extraction of semantic relation between named entities from news articles. Our framework extracts meaningful semantic information regarding these entities that can potentially enrich the image-text relations generated by above mentioned systems.

Our framework automatically discovers semantic topics encoded in news articles based on the group structure of the vocabulary words used in articles. Each word groups defines a semantic topic. Our framework employs multiple techniques to discover such meaningful word groups. We devise a sparse structured logistic regression model for prediction of occurrence of named entities in articles based on the semantic topics discussed in those articles. This modeling scheme can identify the semantic topics which strongly predict occurrence of any named entity. These semantic topics are, in a sense, relevant to that named entity. Relation between two named entities is based on the any semantic topic that is relevant to both the entities. Because of the inherent evolving nature of the news material, such relationships also tend to evolve over time. To account for the evolutionary nature of the news material, we apply our framework to process articles published in different time periods separately. Such application scheme enables the framework to track the existence and the nature of relations between named entities over long periods of time.

In past, the problem of extracting relations between named entities from free-flowing text has been studied under the umbrella of Open Information Extraction (OpenIE)[124, 28, 102]. OpenIE systems are heavily dependent on hand-crafted rules to detect relations between named entities. If an OpenIE system incorporates the capability of learning such rules, it generally needs manually extracted tuples of related named entities to bootstrap the training. Such systems usually identify only a handful of pre-defined relation types. In comparison, our approach is completely unsupervised, does not require hand-crafted rules or seed tuples of related entities, and also puts no restriction on the type of relations being discovered.

Figure 4.10 in Section 4.4 of Chapter 4 shows frequency patterns for various named entities in news articles collected from the website of the Time Magazine¹. Some named entities like ‘Adam Lanza’ are mentioned in news articles very frequently for a very short period of time. The exact

¹www.time.com

time period of frequent mention coincides with some major news event involving such a named entity. For example, ‘Adam Lanza’ was mentioned frequently in news articles when a person of this name opened fire in an elementary school in Newtown, Connecticut, in December 2014. For such entities, it is important to identify their relation to the specific news event, the event type, as well as other entities playing any role in that specific event. Named entities like ‘Barack Obama’ are mentioned frequently in news articles for long periods of time. Relations involving entities like ‘Adam Lanza’ are only valid for short periods of time while the nature of relations for entities like ‘Barack Obama’ changes over time. Hence, the semantic relations between named entities tend to evolve, rather than remaining static over time. It is necessary for the system building semantic relations between named entities to cater to such evolutionary nature of information.

Our time-based application scheme for the framework and the evolutionary nature of the news articles collection allow for tracking of changes in relations between named entities. This evolutionary information is lost when a relatively static database like Wikipedia articles collection is used. Wikipedia database is highly structured where articles are linked to each other through hyper-links. This database has been widely used for establishing links between named entities. Such links do not capture evolving nature of named entities’ relations. We employ this database as a verification tool for the evaluation of our system for extracting semantic relations between named entities.

CHAPTER 2: RELATED WORK

This chapter provides a detailed survey of systems previously proposed to solve the problems defined in Chapter 1. Each section of this chapter deals with the research literature of one particular problem.

2.1 Image Annotation and Caption Generation

The problem of automatic image annotation deals with the challenge of predicting concise and accurate set of words that describe an image. Such a system has vast potential applications for image search, retrieval, and organization systems. Due to huge potential benefits of developing such a system, the problem of image annotation has been studied widely in image processing, computer vision, as well as natural language processing communities. The following is an in-depth survey of various previously proposed approaches for image annotation and caption generation.

Relevance models from the domain of machine translation were adapted to solve the problem of automatic image annotation[12]. In machine translation, the goal of the system is to come up with the best sequence of words in one language that matches the contents of the input word-sequence of another language. Relevance models aim at estimating the joint probability between words of different languages. This joint probability distribution is maximized to generate the output sequence most suitable to the input sequence. Such models assume the availability of training data, i.e., a dataset of corresponding word sequences of both languages. An expectation process over the training data is the basis for joint probability estimation.

In case of image annotation, it is assumed that the visual contents of an image are to be ‘translated’ into words[49, 64, 30]. Joint probability of words and visual contents needs to be estimated.

Words can be easily represented in terms of the binary values indicating their presence or absence. On the other hand, representing the visual contents is far from straightforward. Low-level visual features are described in terms of the mean and the standard deviation of color channels and the energy of edge detection filters. SIFT assigns scale-invariant feature vectors to points-of-interest in images[72]. *Blob*-based representation assign color and texture measures to uniform patches of images[104]. As described in Section 1.1 of Chapter 1, such continuous domain feature vectors can be transformed into discrete domain to represent images as *bag-of-words* (BoW). This discretization process introduces quantization error.

Nearest-neighbor type algorithms have also been employed to generate image descriptions. Such algorithms identify the most similar or the ‘nearest’ training images to the test image. Words associated with these nearest neighbors are propagated to the test image. Some form of iterative optimization is employed to estimate the distance between images such that the likelihood of correct word prediction is maximized. Various schemes for estimating the most effective distance metric have been proposed in the past[39, 99, 18, 68, 120]. Such methods marked a substantial improvement in performance over the relevance model based systems at the cost of computational complexity. The iterative optimization employed by such systems increases the computational cost substantially as compared to the relevance models that require only one-pass over the training data for image annotation. Some efforts were made to study and limit the computational cost of such systems[18].

Object and action recognition tools from the domain of computer vision, are aimed at automatically identifying the objects and the actions in images or videos. Such tools are extremely beneficial for security and surveillance systems. These systems need to process visual feeds from various cameras to detect any unusual occurrence in a timely manner. Such objects and actions are also very important parts of the visual contents of images when it comes to describing the images in words. This idea has been the basis for a large number of image annotation systems proposed in the

past which rely heavily on the object and the action recognition tools. The names of the recognized objects and actions are either directly associated with images, or are further processed to generate sentences, or pick the best description from a collection of available word descriptions[83, 61]. Annotation systems that depend on recognition tools are inherently bound by the performance of these tools. Despite the advances made to meet the challenges of object and action recognition in natural scenes, these tools are still limited in their application in real world unconstrained settings. Annotation systems based on these tools are also capable of only dealing with small vocabulary sets, mainly consisting of names of objects and actions for which recognition tools are available. Naming objects and actions is insufficient to produce meaningful captions for news images.

Blei et al. proposed latent Dirichlet allocation based probabilistic topic modeling process that treats documents in a collection as mixtures of underlying ‘topics’[9]. Topics are defined as a probability distribution over all words in the vocabulary. Such topic modeling found vast applications in the fields of machine learning, text mining, language processing, as well as image processing. When images are represented in BoW form in terms of *visual words*, such modeling is directly applicable to image databases. Variations of the original topic model were proposed to find correspondence between words of two different types, such as textual words and *visual words*[8, 95, 97]. As discussed earlier, BoW representation for images introduce quantization error in the system, undermining the effectiveness of the topic estimation model.

The goal of the caption generation system is to come up with sentence-like descriptions for test images. One way to approach this problem is to find the most appropriate words to be associated with the image and then use these words to generate sentences. Kulkarni et al. employed object and action recognition tools on images, as well as measuring image characteristics such as color. They later used names of objects and actions, and image characteristics as nouns, verbs, and adjectives, respectively, in a language model to generate sentences[61]. To filter the noisy outputs of recognition tools, word co-occurrence statistics were employed[83]. For each query image,

human-composed phrases used to describe visually similar images were collected and selectively combined to form a unique and accurate description[62]. Yang et al. employed the state-of-the-art object and scene detectors to identify object names, i.e., nouns, and scene categories. They later used a language model trained over English Gigaword corpus to estimate the probability of verbs or actions associated with the detected nouns and scenes. They used these estimates as parameters for a hierarchical Markov model (HMM) to generate sentences[126]. Ushiku et al. proposed that the image contents can be described in terms of ‘multi-keyphrases’. Each ‘keyphrases’ describes some prominent image feature. They devised an online learning method to estimate these *keyphrases*. Their proposed framework combines *keyphrases* for each image through an experimental grammar to generate sentence describing prominent image features[118]. An effective annotation method is needed for such methods to work.

An alternative approach is to assume that a large database of image descriptions is available. With this assumption, the problem of caption generation is equivalent to the retrieval problem. The system is provided with an image and it is supposed to retrieve the best caption for this image. Ordonez et al. assumed the availability of large collection of images with their appropriate captions. Their framework searches for the closest matching image in this collection, to the input query image, and transfers its caption to the query image[88]. They concluded that the performance of such a framework improves with the availability of larger collection of image-description pairs. Hodosh et al. framed the problem of caption generation as ranking appropriate hand-written descriptions for images [45]. Gong et al. employed weakly annotated photo collection for image-sentence embedding[38]. Socher et al. developed a system to find appropriate images given a sentence [107]. Kuznetsova et al. attempted to generate an appropriate training database of image-caption pairs by generalizing available captions through a sentence-compression method guided by visual contents of images. They released a large database of image-description pairs in which visual and textual contents are tightly aligned[63].

When caption generation problem is re-framed as a retrieval problem, time-efficient methods can be devised to solve this problem but this approach is extremely limited in its application as it assumes the availability of a suitable database of descriptions. For a test set of news images, it is unreasonable to assume that their appropriate captions can be picked from past news image-caption pairs as news material is supposed to present current events. Discussion about current events may bear similarity to the past event, but is unique by the very nature of news material.

Deep convolutional neural networks (CNN) have gained tremendous popularity for tasks such as hand-written digit recognition and object detection. LeCun et al. proposed a CNN framework for digit recognition[65]. This framework has been widely adapted to deal with the challenge of object detection[59, 106]. For effectively training a CNN for object detection, very large labeled dataset is required. Availability of the ImageNet database solved this problem. ImageNet is a large hierarchical database in which images are placed into classes based on the objects they represent[23]. Feature vectors for images optimized by CNN framework pre-trained over ImageNet database for object detection, are widely used by annotation and caption generation systems. Since a properly formed sentence-like image caption is sequential in nature, caption generation frameworks involve some form of sequential neural network such as recurrent neural network (RNN) or long short-term memory network (LSTM) as well[79, 25, 121, 125, 56, 54]. There has been some effort to properly identify and ‘attend’ to the appropriate part of the image while generating its corresponding description. Such frameworks are called *attention models*[125]. These deep neural network based image description systems have enjoyed tremendous success in recent past.

Despite the tremendous popularity of neural network based models for caption generation, there are certain disadvantages of such models. These models employ image feature vectors optimized by CNN framework trained over the ImageNet database for object detection[59, 106]. These models have been extremely successful when tested over social media images showing people engaged in everyday activities. MSCOCO is a very large database of such images[70]. Ground truth cap-

tions for its images have been collected through crowd-sourcing. Human caption writers were asked to write sentences to describe image contents. Five captions were collected for each image. These captions describe image contents in terms of the objects present in the image or simple actions taking place that involve some objects. Nature of such image captions is similar to the standard image annotation benchmark datasets such as IAPR TC-12 and ESP game. Visual features generated by ImageNet-trained CNN for such images correspond to the common objects that are part of ImageNet labels set, and are also named in ground truth image captions. In other words, CNN-based visual features correspond closely to ground truth image captions as these captions contain words similar to the ImageNet labels. Real world image captions do not simply describe the objects presented in images. They hint at the *context* of the image. People uploading photos of vacations or parties describe those images in reference to the occasion those photos were taken at. *Semantic gap* becomes even wider in this case. Inclusion of such *context* in the image description seems beyond the scope of CNN-extracted visual features.

We deal with the inherent *semantic gap* between visual and textual contents by incorporating semantic contextual information from every source possible. There is no direct correlation between simple visual and textual representation units. Instead of focusing on complex and time-consuming strategies to devise new representation schemes, we devise a framework to extract and quantify contextual information from the dataset. Data items can be grouped in such a way that each group defines a semantic topic. Association between a new data item and such semantic groups encode the semantic contextual information of the new item. The next step in this process is to incorporate these contextual cues when establishing inter-modality relation, e.g., image-word association in the process of image annotation. We devised various strategies for extracting semantic contextual relations and employing them in annotation models. Chapter 3 describes details of our strategies. In Chapter 4, we introduce *context* extraction strategies to deal with information sources of various data types, in the context of automatic generation of news image descriptions. We thoroughly

evaluated our frameworks against previously proposed framework. We can safely conclude that the incorporation of the *context* is an effective and efficient way to generate suitable image annotations.

2.1.1 News Image Annotation and Caption generation

As explained in Section 4.2.1, ‘artificial’ descriptions of images in datasets like MSCOCO and Flickr30K are used as ground truth to test image annotation systems, completely ignoring their real world captions which include hints to the *context* of images as well as their contents. News images and their real world captions are commonly available on websites of news media outlets. The *semantic gap* between such images and their real world captions seems much wider than the gap between images and their artificial descriptions in standard image annotation datasets. Nonetheless, image annotation systems need to be able to match real world captions for their effective incorporation into any practical search, retrieval or organizational system.

The problem of news images annotation has not been studied as widely in the past, as the traditional image annotation problem. There are a few possible approaches to deal with this problem. One approach is to use any image annotation system designed for standard image annotation datasets to produce annotations for news images. This way, contextual or auxiliary information available with the image such as news article or metadata related to the article, are completely ignored. Sample news image-captions pairs in Figure 4.4 in Section 4.2 of Chapter 4 show that the images with similar visual contents may have different captions because of the different *context* described in their corresponding articles.

One argument is that if image is already ‘annotated’ with large amount of text in some form, e.g., news article, the information contained in that text is enough to predict annotations for the image. Some previously proposed methods rely only on knowledge mining through this annotated text to predict annotations[66, 67, 80, 20]. Leong et al. devised various text mining features to extract

keywords or appropriate annotations for images, from the text surrounding those images on their respective web-pages[66, 67]. They also discussed how some of the words are more ‘picturable’ than others, hence more likely to be used as image annotation[66]. Choi et al. produced annotations for news images by semantically analyzing their associated text in the form of articles and article titles[20]. Mason et al. discussed some baseline methods for evaluation of image annotation for image databases with associated text[80]. Such methods completely ignore the information contained in visual contents of images. Image-caption pairs presented in Figure 4.5 in Section 4.2 of Chapter 4 shows that images may have different captions, even when their corresponding news articles discuss the same topics, because of their visual contents.

Feng et al. proposed a framework that combines information from both visual contents of images and textual contents of their accompanying news articles[33, 31, 32]. They proposed a framework that extended the application of relevance model based annotation system to incorporate the influence of the text of accompanying news article. This influence was essentially incorporated through re-ordering the words selected as annotations based on joint probability estimated by relevance model[31]. They later proposed another model that employed BoW image representation and latent Dirichlet allocation based topic modeling to find correlation between visual and textual words[32]. Given the quantization error introduced through BoW image representation, estimated correlation between images and text is only moderately effective. They collected their own dataset of images to evaluate their approach. This dataset is described in Section 4.1 of Chapter 4. Vocabulary size for such news image collections is very large because they include large amount of text in the form of news articles. Feng et al. used a substantially reduced vocabulary set to report results[33, 80].

Feng et al. also proposed complex models to generate sentence-like captions for news images. Their models include language modeling to assign probability values to word or phrase sequences, as well as joint probability of visual and textual contents estimated through annotation framework.

Language modeling is supposed to enforce the grammatical correctness to generate meaningful sentences. Their reported results show that the generated captions are not often meaningful[33].

We devise a framework that can efficiently and effectively predict annotations for news images and can also pick the best sentences to describe these images. This framework incorporates information from not only images and articles, but every resource available. We generalize the semantic information extraction schemes that we present for standard image annotation (described in Chapter 3), to include a variety of data types by transforming semantic information from all sources into a common representation space. We choose *probability space* as the common representation space. Our model design is focused on dealing with large vocabulary sets while being efficient in terms of the time and the computational complexity. We employ a *context*-sensitive generative model inspired by relevance models for predicting annotations for news images. We also devise an *extractive* framework to associate images with their best sentence-like caption, based on information collected from article as well as predicted annotations for images. Our strategies for annotation and caption generation for news images are described in detail in Sections 4.2 and 4.3 of Chapter 4, respectively. Through evaluation of these frameworks prove the merits of our *context*-sensitive modeling of the problem of automatic generation of realistic image captions, over various types of previously proposed systems.

2.2 Semantic Network of Named Entities

As explained in Section 1.2 of Chapter 1, image-text relations for collections of news items can be further enriched through the exploration of semantic relational information among linguistic features. Linguistic features are words used in free-flowing text like news articles which have special meaning. Named entities, i.e., the words indicating the names of people, places and organizations, are the most important linguistic features for news articles and blogs as they constitute the

most-often used search queries[82]. Such entities are also important in reference to the image-text relations as there have been many attempts to link such entities directly to images. Recognition of people in news images have been studied in [5, 89]. Such systems directly link entities of the type ‘person’ with images. Identification of landmarks from web images have been explored in [44, 1, 17], linking entities of the type ‘place’ with images. Wide variety of systems have been proposed to identify logos and trademarks in natural scenes, associating images with entities of the type ‘organization’[29, 55]. We devise a framework to automatically extract meaningful semantic relations between named entities through the exploration of semantic topics discussed in news articles.

The idea of discovering and understanding the links between the named entities through text analysis has been widely used in semantic web and natural language processing communities. The most basic task is to extract and identify the named entities and their types from free-flowing text. Stanford NER is a very popular named entity recognition tool developed by the Stanford natural Language Processing Group¹. This tool employs conditional random fields modeling to identify the sequences of words that constitute named entities in text.

Entity-linking tools aim at linking named entities mentioned in the text to their corresponding entries in some database containing information about known named entities[50, 51, 103]. Wikipedia² is an extensive structured database of vast information about named entities. When a system links a named entity mentioned in the text, to its Wikipedia page, all the information regarding this entity and its association with other entities become accessible to the system. Deeper understanding of the text can be developed using this additional information. Ambiguity and inconsistency in the way named entities are mentioned in the text, are the main challenges for any named entity linking tool. Jin et al. argued that only a handful of lexical features, out of all lexical features of the text, are

¹<http://nlp.stanford.edu/software/CRF-NER.shtml>

²www.wikipedia.com

crucial to the semantic resolution of named entities. Hence this problem can be modeled as sparse signal recovery problem[51]. They later proposed another sparse signal processing based model to correctly link unpopular or infrequently mentioned named entities[50]. Shen et al. employed both Wikipedia and rich semantic knowledge embedded in WordNet for effective named entity linking. WordNet is a large database of English nouns, verbs, adjectives and adverbs which are grouped into synonym sets³. Links between these sets encode semantic and lexical information.

Traditionally, unsupervised information extraction systems were built to harness the power of search engines to extract knowledge from web documents. KNOWITALL system is given an extensible ontology and some domain-independent extraction templates. It creates text extraction rules for each class and populates the ontology with information about the named entities using these extraction rules as search queries for search engines[27]. Open information extraction (OpenIE) systems seek to broaden the scope of unsupervised information extraction, to extract large number of facts from text document collections without human supervision. TEXTRUNNER tool was based on self-supervised learning. The tool first trains a classifier over a set of *trustworthy* and *untrustworthy* relational tuples of named entities. Later, lightweight noun phrase parsers are used to extract tuples of named entities mentioned in sentences together and the learned classifier is used to judge their quality[26]. Many OpenIE tools inherently rely on named entities to be mentioned together in sentences of specific structures to extract relations between them. A common restriction for the sentence structure is to connect two entities through a verb[124, 28]. WOE tool incorporates similarity between infobox⁴ entries of Wikipedia pages of named entities, to substantiate relations between them[124]. OLLIE alleviates the restriction of sentences to involve a verb, but requires the output of ReVerb[28] to be provided as input that employs the same restriction[102].

³<https://wordnet.princeton.edu/>

⁴Infobox on Wikipedia page contains a few tuples summarizing the characteristics of the named entity that is the subject of that Wikipedia page. For example, infobox on Wikipedia page of University of Central Florida contains attributes like ‘Location’, ‘President’, ‘Established’, ‘Mascot’, etc.

Typically, systems for building semantic networks of named entities are either supervised or semi-supervised [11, 2, 130, 81, 24, 43]. Such systems often require an initial seed in the form of pairs of named entities with a known relation. These systems aim at finding other pairs of named entities with the same type of relation between them[11, 81]. Freebase⁵ is a database of well-known people, places and things. Hence, it is an excellent source of relational information about named entities and is often used by systems to get the seed relational tuples of named entities[81]. The scope of such supervised or semi-supervised relation extraction systems is rather limited. Such systems are restricted to discovering relations of a limited variety and may also require external databases like Freebase and Wikipedia for their working. Less work has been reported on unsupervised methods for relation extraction. Rosenfeld et al. propose named entity clustering such that the named entities appearing in the same context are put in one cluster [98]. Hasegawa et al. propose unsupervised relation discovery among named entities appearing in the same sentence[42]. These systems are unsupervised, but impose restrictions on named entities to be considered for relation discovery, and also assign only a handful of manually picked labels to discovered relations.

Most of the previously proposed systems are aimed at building a knowledge-base of facts [11, 2, 130, 81, 24, 43, 102, 28]. Therefore, these systems extract relations based on a limited number of linguistic patterns connecting named entities in individual sentences. In comparison, we devise a framework that models named entities' occurrence through sparse structured logistic regression model. Structure among the predictors, i.e., the vocabulary words, indicate semantic concepts or topics. Coefficients estimated from sparse structured modeling indicate the semantic topics which correlate closely with named entities. Semantic relations between two named entities can be based on their common relevant topics. We devise multiple strategies to automatically extract semantic concepts from news articles.

⁵www.freebase.com

Our framework overcomes various limitations of previously proposed systems such as the availability of a seed set of related named entities, hand-crafting of extraction rules, discovering relations between entities mentioned within a sentence only, use of external database. Our system has vast potential application in news search and retrieval systems as well as news recommender tools. Such tools should direct readers to articles of their interest based on articles they are currently reading. The semantic relations discovered by this system can greatly enrich the image-text relations identified by the systems that link images to named entities like people, logos, landmarks, etc. The scope and nature of our system are largely different from those of OpenIE systems which are focused on building databases of general knowledge.

CHAPTER 3: AUTOMATIC IMAGE ANNOTATION

In this chapter, we describe various systems that we developed for automatic annotation of images with appropriate words. All the systems discussed in this chapter are focused on annotating images available in standard image annotation datasets with no auxiliary information sources available. The core idea in each of these systems is to develop contextual relations between images and semantic concepts or topics. Since no auxiliary information sources are available, semantic topics are extracted from image-description pairs available in the training dataset. These semantic relations are then incorporated in the system searching for appropriate words for images as prior knowledge. We devised various strategies for extraction of semantic concepts or *themes*. We also explored different techniques for understanding the semantic relations between the images and these concepts, as well as modeling schemes for predicting annotations for images in reference to their semantic relations.

Some common notations will be used in each of the proposed models. The automatic image annotation system is provided with an image Y and the system is expected to return a set of words ($\mathbf{w}_Y = \{w_{y1}, w_{y2}, \dots, w_{yB}\}$) such that each word $w_{yb} \in \mathbf{w}_Y$ is an appropriate tag for the image Y . System is also supplied with a training set consisting of labeled images or image-description pairs. Vocabulary set consists of the words used in the descriptions of training images. Let \mathcal{X} and \mathcal{W} denote the sets of training image-description pairs and vocabulary. Let M and N denote the sizes of set \mathcal{X} and \mathcal{W} , respectively. Our work is aimed at extracting semantic topics or *themes* from the available training dataset. These topics or *themes* are defined in terms of groups of training items. Each of the proposed system devises a framework to split the training data \mathcal{X} into non-overlapping clusters such that all image of k^{th} cluster C_k represent a certain semantic topic or *theme* in their visual contents and textual descriptions. Let \mathcal{C} denote the set of clusters or *themes* such that K is the number of clusters, i.e., $K = |\mathcal{C}|$.

3.1 Scene-based Automatic Image Annotation

Studies have shown that scene identification in humans is independent of individual object identification. Humans can integrate enough information about a scene in less than 200 ms. The ‘gist’ of the scene is identified as quickly as identification of a single object[94, 6, 86]. Scene recognition is about broad understanding of semantic properties of visual contents of an image while the appearance of objects in images constitute the details of the image. We devised a system to harness the semantic understanding of visual contents of images in terms of scene identification to bridge the *semantic gap* between images and words.

Efficient mathematical description of scene information of images was a big question for our work. For our first image annotation model, we turned our attention to the work of Oliva et al[86]. Oliva et al. also based their work on scene description on the hypothesis that scene identification relies on the global characteristics of images while being independent of the local image segments. They proposed a set of perceptual dimension, i.e., openness, roughness, naturalness, ruggedness, expansion, etc. These perceptual dimensions represent dominant spatial structure of scenes. They employed both Fourier transform and Principal Component Analysis (PCA) to estimate the perceptual characteristics of images in terms of their spectral properties. They showed that the second order statistics of images are constrained by their scene categories. Each images is projected onto the perceptual dimensions and the resulting feature vector is called *spatial envelop* of the image. In this dissertation, we refer to the image feature vectors obtained through the method described by Oliva et al. in [86] as GIST. The GIST feature vector for any image is an efficient and effective description of its semantic scene category.

Since scene represents global semantic properties of images, we employed GIST representation of images to define semantic concept in terms of scene-categories. Each scene-category encodes some information about the appearance of the objects and other characteristics of images. Figures

3.1 and 3.2 show sample images for two scene-categories. If an image belongs to scene-category of Figure 3.1, it is highly likely that the words like ‘face’ and ‘child’ are associated with this image. Similarly, images belonging to category of Figure 3.2 are likely associated with words like ‘people’, ‘mountain’ and ‘grass’. Hence, the association of any image with these semantic scene-categories can provide highly valuable information to the system predicting word annotations for images. In the following section, we describe ‘scene-AIA’, i.e., the system that we devised to incorporate semantic scene characteristics in the process of image annotation. This work was published in IEEE International Conference on Image Processing, 2014[111].

3.1.1 System Architecture

We assume that each image is made up of A number of visual units i.e. $\mathbf{r} = \{r_1, r_2, \dots, r_A\}$. As explained in Chapter 1, the selection of visual feature description scheme is both challenging and important for image annotation models. To avoid the quantization error induced by BoW visual feature description approach, we employed a grid based continuous domain visual representation scheme. Each image is divided by a fixed grid. Feature vector describing the color and the texture properties of one grid section is the smallest visual unit. Section 3.1.2.2 describes how color and texture characteristics of the image are described in each visual unit.

Description of each training image is assumed to be made up of B number of words $\mathbf{w}_X = \{w_{x1}, w_{x2}, \dots, w_{xB}\}$. We assume that there are certain scene-categories available, contained in the set $\mathcal{C} = \{C_1, C_2, \dots, C_K\}$. Each scene-category is defined by a group of training images which display a certain type of scene. When a test image Y is provided to the system, its association with available scene-categories is estimated. Test image Y may display characteristics of multiple scene-categories. Therefore, we do not assume that Y belongs to one scene-category only. Rather, the association of the test image Y with all available scene-categories is encoded in a probability distribution $P(\mathcal{C}|\theta_Y)$ where variable θ_Y contains information about the scene characteristics of Y .

While predicting appropriate words for image Y , the system takes the distribution $P(\mathcal{C}|\theta_Y)$ into account.

3.1.1.1 Scene Categorization

We compute GIST features for all training images and cluster them through hierarchical clustering based on cosine similarity between GIST features of images. This clustering scheme uses maximum allowed size of the cluster as a system parameter. Any cluster that is larger than the maximum allowed size, is further divided by hierarchical clustering. Clusters with very small number of members are also dropped. The goal is to come up with image clusters representing scene-categories such that the size of any cluster falls within a narrow range. This restriction ensures that the training data is relatively evenly distributed among scene-categories, and the training process does not unduly favor any scene-category.

Let us assume that this clustering process generates K sets of images such that each set \mathcal{X}_k of size M_k , corresponds to one semantic scene-category C_k . For training image $X \in \mathcal{X}$

$$P(X|C_k) = \begin{cases} 1/M_k, & \text{if } X \in \mathcal{X}_k. \\ 0, & \text{otherwise.} \end{cases} \quad (3.1)$$

K can be selected using a threshold on within cluster entropy. We observed that the performance of our system remains stable for a wide range of K and reported the best results. Figures 3.1 and 3.2 show sample images for two such clusters for IAPR TC-12 dataset.

We compute GIST representation for every test image Y , i.e., the image whose word annotations need to be predicted. This representation, denoted by θ_Y , encodes the information about semantic scene characteristics of image Y . Association between Y and each scene-type is quantified in

the form of the following conditional probability distribution involving a non-parametric Gaussian kernel.

$$P(C_k|\theta_Y) = \frac{\exp(-(G_{X_k}^Y - \theta_Y)^T \Lambda^{-1} (G_{X_k}^Y - \theta_Y))}{\sqrt{2\pi|\Gamma|}} \quad (3.2)$$

$G_{X_k}^Y$ is the GIST representation of the member of cluster \mathcal{X}_k corresponding to scene-category C_k , which is the closest match to θ_Y of the test image at hand. Λ is the covariance matrix, assumed to be of the form $\kappa \mathbf{I}$ where \mathbf{I} is the identity matrix and κ can be selected empirically over held-out data.



Figure 3.1: A sample *Scene* from IAPR TC-12 dataset.



Figure 3.2: A sample *Scene* from IAPR TC-12 dataset.

We also add a ‘general’ category to the set of scene-categories \mathcal{C} and assume that the cluster for this category consists of all training images. Idea is that many words are specific to some scene-category but some words are generic and appear in descriptions of images presenting varying types of scenes. Processing for the type ‘general’ provides evidence for those words. $P(C_k|\theta_Y)$ where C_k represents the ‘general’ category, is assigned a fixed weight for all images. Overall $P(\mathcal{C}|\theta_Y)$ is then renormalized. The fixed weight can be determined over a held out portion of the data.

Distribution $P(\mathcal{C}|\theta_Y)$ is the continuous domain representation of semantic information of test image in terms of its associations with all semantic scene-categories. This distribution emphasizes the fact that the test image may show characteristics of multiple scene-types and should be annotated in the light of its association with all semantic categories.

3.1.1.2 Relevance Model based Image Annotation

Our annotation model is inspired by the relevance models from the domain of machine translation. This annotation model estimates joint probability distribution of visual and textual representations of images, conditioned over the semantic information of a given test image denoted by θ_Y . Hence, the proposed model is sensitive to the semantic *context* of the image. The following expectation process is employed to estimate such joint probability distribution.

1. pick a scene-category $C_k \in \mathcal{C}$ with probability conditioned over variable θ_Y i.e. $P(C_k|\theta_Y)$
2. pick image X from training set \mathcal{X} with probability $P(X|C_k)$
3. for $a = 1, 2, \dots, A$
 - (a) pick a visual unit r_a from conditional probability $P_{\mathcal{R}}(\cdot|X)$
4. for $b = 1, 2, \dots, B$
 - (a) pick a word w_b from conditional probability $P_{W_{C_k}}(\cdot|X)$

Thus, joint probability of \mathbf{r} and \mathbf{w} conditioned over θ_Y is given by the following equation.

$$P(\mathbf{w}, \mathbf{r}|\theta_Y) = \sum_{C_k \in \mathcal{C}} P(C_k|\theta_Y) \sum_{X \in \mathcal{X}} P(X|C_k) \prod_{b \in B} P_{W_{C_k}}(w_b|X) \prod_{a \in A} P_{\mathcal{R}}(r_a|X) \quad (3.3)$$

We used multinomial distribution for modeling image descriptions. Thus, $P_{\mathcal{W}_{C_k}}(w_b|X)$ is the w_b^{th} component of multinomial distribution over the words in the set \mathcal{W}_{C_k} which generated the description for sample X of the training data. \mathcal{W}_{C_k} is the vocabulary for samples of the scene-category C_k . Bayes estimation for this distribution, given beta prior is given by the following formula.

$$P_{\mathcal{W}_{C_k}}(w_b|X) = \frac{\mu\delta_{w_b} + M_{w_bk}}{\mu + M_k} \quad (3.4)$$

M_k is the size of the set \mathcal{X}_k corresponding to scene category C_k . M_{w_bk} is the number of samples from the set \mathcal{X}_k with the word w_b in their ground truth descriptions. δ_{w_b} is 1 only when the ground truth annotations for the training image X contain the word w_b . Constant μ can be selected empirically over a held-out portion of data.

$P_{\mathcal{R}}(r_a|X)$ is the density estimate for generating visual unit r_a given a training image X . We used a non-parametric Gaussian kernel to estimate this density. Assuming that the training image X consists of A visual units $\mathbf{r}_X = \{r_{x1}, r_{x2}, \dots, r_{xA}\}$

$$P_{\mathcal{R}}(r_a|X) = \frac{\exp(-(r_a - r_{xa})^T \Sigma^{-1} (r_a - r_{xa}))}{\sqrt{2\pi|\Sigma|}} \quad (3.5)$$

The covariance matrix Σ is assumed to be of the form $\beta\mathbf{I}$ for convenience where \mathbf{I} is the identity matrix. Variable β determines the smoothness around the point r_{xa} and can be empirically selected over a held-out portion of the dataset. Note that this kernel signifies the importance of the spatial coherence while quantifying the similarity between two images as it compares visual units at corresponding grid positions only. If index ‘ a ’ represents some information other than the position of the grid section, e.g., the type of visual feature, this kernel would still be able to correctly quantify similarity by comparing units of similar type with each other. This property is important to generalize this annotation model over visual features other than the grid-based representation scheme.

3.1.2 Evaluation

We thoroughly tested two variations of our framework over two different datasets. The following is the detail of our evaluation scheme as well as experimental results and their implications.

Note that the image annotation systems are used to produce as many annotations per image as is the average number of words associated with images in the training data. Mean values of the precision and the recall per word and the number of words with positive recall (N^+) are reported as performance evaluation measures.

3.1.2.1 Datasets

We evaluated our system on two popular image annotation datasets, i.e., IAPR TC-12¹ and ESP². IAPR TC-12 dataset contains 19,846 images, each described carefully in a few sentences. Frequently occurring nouns, verbs and adjective, are picked to form the vocabulary set after tokenizing and part-of-speech tagging these sentences. ESP game dataset consists of images labeled by the players of ESP game. A smaller subset of size 21,844 has been popularly used to test image annotation systems. We used the same split of data in the training and the test sets (90% for training, 10% for test) for both of the datasets as used by other image annotation systems. IAPR TC-12 and ESP datasets have been generally tested over vocabulary sets of 291 and 269 most frequently occurring words respectively, by various image annotation systems. In our system, the vocabulary varies from the set of one scene-category to the other, instead of being fixed to a specific number for all of the dataset. But we made sure that approximately the same number of unique words appear in the final output, i.e., the annotations predicted for test images, by adjusting the param-

¹<http://www.imageclef.org/photodata>

²www.espgame.org

eters of our system. We report the results over these unique words to keep them comparable to those of other systems. We used approximately 50 scene-categories clusters for both datasets after dropping too small clusters.

3.1.2.2 *Visual Features*

We used 5×6 grid to divide images and assigned each grid section a feature vector of length 46, representing its color and texture characteristics. This representation scheme is the same as used by many other image annotation systems, e.g., [64, 30]. Each feature vector contains 18 color features (mean and standard deviation of each channel of RGB, LUV and LAB color-spaces), 12 texture features (Gabor energy computed over 3 scales and 4 orientations), 4 bin histogram-of-gradients (HoG) and discrete cosine transform coefficients. We observed that increasing the grid size beyond 5×6 did not improve performance.

Guillaumin et al. observed an improvement in the performance by using a combination of holistic and local visual features[39]. More recently, Chen et al.[18] and Verma et al.[120] used the same features in their systems. We also employed these features and observed an improvement in the performance of our annotation model.

3.1.2.3 *Results*

Scene-AIA represents our system employing the grid-based visual features described in Section 3.1.2.2. **Scene-AIA-B** represents our system making use of the visual features devised by Guillaumin et al. [39]. Tables 3.1 and 3.2 show performance comparison of our annotation model against various previously proposed annotation systems over two datasets, i.e., IAPR TC-12 and the ESP datasets, respectively.

Table 3.1: Performance evaluation for IAPR TC-12 dataset

Model	Mean precision per word	Mean recall per word	N ⁺
CRM[64]	21	15	214
MBRM[30]	21	14	186
MBRM-G[39]	24	23	223
BS-CRM[84]	22	24	250
JEC[77]	25	16	196
Lasso[77]	26	16	199
HGDM [69]	29	18	–
AP[99]	28	26	–
TagProp-ML[39]	48	25	227
TagProp[39]	46	35	266
FastTag[18]	47	26	280
2PKNN-ML[120]	54	37	278
Scene-AIA	55	20	254
Scene-AIA-B	56	25	230

Table 3.2: Performance evaluation for ESP game dataset

Model	Mean precision per word	Mean recall per word	N ⁺
CRM[64]	29	19	227
MBRM[30]	21	17	218
MBRM-G[39]	18	19	209
JEC[77]	23	19	227
Lasso[77]	22	18	225
AP[99]	24	24	–
TagProp-ML[39]	49	20	213
TagProp[39]	39	27	239
FastTag[18]	46	22	247
2PKNN-ML[120]	53	27	252
Scene-AIA	45	19	246
Scene-AIA-B	60	20	234

Our system outperforms other generative probability estimation based methods such as CRM[64], MBRM[30], BS-CRM[84]. it also outperforms systems based on greedy algorithms like JEC and Lasso[77]. MBRM-G denotes the case when MBRM[30] model employs visual features defined by Guillaumin et al.[39]. These visual features include the GIST representation of images. Our system

performs much better than MBRM-G implying that our semantic *context*-sensitive modeling is a more effective way of incorporating scene information in the annotation process.

Nearest-neighbor based approaches such as TagProp[39] and 2PKNN-ML[120] employ iterative optimization algorithms, rendering them computationally expensive and not particularly scalable to larger datasets. Chen et al. proposed FastTag to reduce the computational complexity and presented a detailed complexity analysis of different annotation systems[18]. Our system is computationally efficient as it employs a generative modeling scheme that needs only one-pass over the training data to estimate the joint probability of words and visual features. Clustering based on scene-categories is only required for training sample and can be pre-computed using efficient clustering algorithms. Efficient clustering algorithms are practically less time-consuming than iterative optimization algorithms used by TagProp[39] or 2PKNN-ML[120] because of their better termination conditions. The system still outperforms TagProp, FastTag and 2KPNN-ML in terms of the mean precision and is comparable in terms of the mean recall against FastTag and TagProp-ML.

3.1.2.4 Cluster Expansion for Large Datasets

To prove the scalability of our system, we tested it over complete complete ESP dataset, referred to as the ESP-large in this document. This dataset contains 67796 image-description pairs (90% dataset for training, 10% for testing). We used grid-based visual features and reported results over a set of 1400 unique words. We generated roughly 200 clusters based on scene-categories using efficient implementation of K-means clustering³. Our assumption is that the larger dataset contains greater variety of scene characteristics. Hence, they require larger number of scene-category based clusters to effectively encode the variation in their semantic scene properties. We compared the performance of our system against MBRM. MBRM[30] employs generative modeling for prob-

³<http://www.vlfeat.org/>

ability estimation and is computationally very efficient. Table 3.3 shows that our system beats MBRM for ESP-large dataset; proving that our system is scalable for larger datasets with vast vocabulary sets.

Table 3.3: Performance evaluation for ESP-large dataset

Model	Mean precision per word	Mean recall per word	N ⁺
MBRM	34	15	770
Scene-AIA	47	24	979
Scene-AIA-exp	44	23	901

We tried another variant of our system, named **Scene-AIA-exp** to reduce the computational complexity even further. We split the training data in two halves and used the clustering algorithm on one half of the training data. Then we expanded the clusters by adding each image from the other half of the training data to the cluster containing its closest match based on the GIST features. Thus, even the computationally efficient clustering algorithm needs to be run over only half of the training data. Only slight reduction in the performance is observed. This also indicates that our system is flexible enough to make use of additional training data as it becomes available without having to restart the training process from scratch.

Thorough evaluation of our scene-based image annotation model scene-AIA clearly indicates that the semantic scene properties of images encodes valuable semantic information regarding the contents of images. Image annotation models are generally focused on identifying image contents so that they can be mentioned as textual annotations. We devised an effective semantic *context*-sensitive model for employing scene-analysis based semantic information during the process of predicting annotations for individual images. Our semantic *context*-sensitive model does not only outperform various previously proposed systems but also achieves such performance in computationally efficient manner.

3.2 Feature-Independent Semantic Relations Extraction for Image Annotation

Image annotation systems are generally provided with training data consisting of image-description pairs. In addition to scene characteristics of images, their textual descriptions also hint at their semantic characteristics. We devised a strategy to quantify this semantic information in terms of word-groups such that each word group defines one semantic topics or concept. The association between images and these semantic topics is estimated through a feature-independent framework, i.e., no visual features are extracted from images for estimating their semantic properties but raw images are used directly. The estimation process involves tensor analysis of images. Semantic information extracted in this manner is incorporated in a generative model to predict annotations for images. We named this annotation scheme ‘Tucker-AIA’, and it was published in IEEE Conference on Computer Vision and Pattern Recognition, 2015[112].

Tensors have been used as a natural representation scheme for videos, text document collections and image ensembles[21, 60, 119, 3, 41]. Tensor analysis and decomposition algorithms have been used on videos in action recognition and motion detection systems[90, 73, 128, 108]. We devised a unique strategy to build a tensor from individual images of the same semantic topic or *theme*. The tensor is decomposed to generate a *signature* for that semantic *theme*. The relations between test images and these *theme* signatures are also estimated through tensor decomposition. Kolda et al. presented a detailed study of the tensor decomposition methods along with their applications[58]. Our framework employs Tucker decomposition presented in [58].

3.2.1 System Architecture

In this section, we present our framework for automatic extraction of semantic relations of images as well as our annotation model that incorporates these semantic relations.

3.2.1.1 Semantic Relation Extraction Framework

The quantification of semantic relations between images and semantic *themes* is a three-stage process. These stages are; 1) the identification of semantic *themes*, 2) the estimation of visual statistical *signature* for each *theme*, and 3) the estimation of the association between test images and semantic *themes*. The following is a detailed description of all of these stages.

3.2.1.1.1 Semantic Concept-based Categories

Semantic groups of images need to be constructed such that *a)* each image group is a representation of some semantic concept or *theme* that is capable of aiding the prediction of appropriate annotations for images, *b)* images of one group have sufficient visual similarity to each other so that the group can be used as the basis for the formation of visual *signature* of the semantic *theme*.



Figure 3.3: An example of semantic theme formed on the basis of similarity in textual descriptions



Figure 3.4: An example of semantic theme formed on the basis of similarity in textual descriptions

The textual descriptions of images in the training set are available. It is intuitive to assume that the textual description of an image predicts its visual contents. We turned to the literature dealing

with the problem of text search and retrieval. Text search engines employ *tfIdf* representation scheme for text documents and text queries. We decided to employ *tfIdf* representation of image descriptions. Such representation scheme assigns higher weights, and hence more importance, to the *distinctive* words[117]. Words which are too common do not have much information content. Whereas the moderately frequent words or the words which are part of a few textual items only, represent the *distinctive* properties of those items. In case of image-description pairs, words which occur with too many images (e.g., ‘sky’) cannot distinguish one image from the other. On the other hand, words like ‘snow’ or ‘cricket’ belong to only a handful of images, and hence define very *distinctive* characteristics of such images. Our strategy is aimed at defining a semantic *theme* as a group of *distinctive* words.

If \mathbf{v}_X denotes the *tfIdf* representation of the description of image X , it is a vector of length N (the size of the vocabulary set) and each of its entries is defined as

$$\mathbf{v}_{Xn} = \frac{B_{Xn}}{M_{w_n}} \quad (3.6)$$

where B_{Xn} is the number of times n_{th} word appears in the description of image X and M_{w_n} is the number of image descriptions in the dataset that contain n_{th} word. The images in the training set are clustered based on the cosine similarity between their *tfIdf* vectors. The properties of *tfIdf* representation ensures that this process groups images with the same *distinctive* words in their descriptions, together. An image group T_c will be able to uniquely provide evidence for the *distinctive* words shared between its images. We employed agglomerative hierarchical clustering method with cut-off threshold. Very large clusters are further split by the same technique such that the size of each cluster falls within a narrow range. The goal is to achieve relatively even distribution of semantic categories in the training data so that no semantic category is unduly favored during the training phase.

3.2.1.1.2 Semantic Signature through Tensor Analysis

One *semantic tensor* $\mathbb{T}_c \in \mathbb{R}^{G \times H \times J}$ is constructed for each image group T_c formed during the first step. The images in one group are resized to a fixed height H and width G , converted to gray-scale, processed through a Gaussian blurring filter and concatenated together to form the tensor. Three dimensions, i.e., g , h and j , of this tensor represent image width, image height and image indices, respectively.

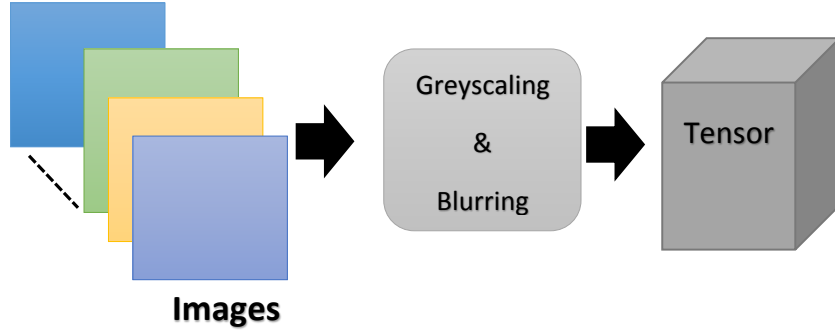


Figure 3.5: Tensor formation: images of one group are stacked together to form one tensor

Notice that the goal of this process is to estimate an overall signature of the semantic *theme* while the *theme* is encoded in the *distinctive* words of image descriptions in one group. This *semantic signature* should be made insensitive to fine visual details. When association of a new image to any of the *context signatures* is assessed, it focuses on global similarity between the new image and the member images of that *semantic group* and not on the local details of images. Therefore, images of one group are all processed by a blurring Gaussian filter to remove sharp distinctions because of edges.

The next step is the decomposition of the *semantic tensor* through Tucker decomposition, to find a

compact *signature* of the *semantic group*. Tucker decomposition is a popular technique to project tensor $\mathbb{T}_c \in \mathbb{R}^{G \times H \times J}$ onto a smaller core tensor S and three matrices P, Q , and R such that

$$\mathbb{T}_c \approx S \times_1 P \times_2 Q \times_3 R = \sum_{g=1}^G \sum_{h=1}^H \sum_{j=1}^J s_{ghj} p_g \circ q_h \circ r_j, \quad (3.7)$$

where $P \in \mathbb{R}^{G \times F}$, $Q \in \mathbb{R}^{H \times F}$, and $R \in \mathbb{R}^{J \times F}$ are the orthogonal matrices, $S \in \mathbb{R}^{F \times F \times F}$ is the core tensor and $F \leq \min(G, H, J)$. The $\bar{\times}_i$ operator denotes the multiplication between a tensor and a vector in mode- i of that tensor, whose result is also a tensor, namely, $\mathbb{A} = \mathbb{B} \bar{\times}_i \alpha \iff (\mathbb{A})_{jkl} = \sum_{i=1}^I \mathbb{B}_{ijk} \alpha_i$.

Rank-1 Tucker decomposition is applied, i.e., F is set to 1. In this case, P, Q and R are the vectors with lengths equal to the width of the image, the height of the image and the size of the semantic group, respectively. Vector $R \in \mathbb{R}^{J \times 1}$ is the most important for our system. This vector represents the similarity/dissimilarity of one image to its neighboring images in the tensor \mathbb{T}_c . All images concatenated together belong to one semantic groups, i.e., they are all visually similar as they all have highly similar textual descriptions. There should only be small variations in the entries of this vector. Vector R is the compact *signature* for the semantic group.

3.2.1.1.3 Semantic Relations through Tensor Analysis

The next step is to quantify the semantic relations of test images in terms of their association with different *semantic signatures*. Let each test image be represented as Y . There is no textual description available for Y . As we explained earlier, *semantic signature* is a vector of length R with little variation across its entries as it is the result of Tucker decomposition of a tensor made up of R visually similar images belonging to one semantic group. If a foreign entity, e.g., a test image Y , is inserted into this tensor at any location, say l , it will disturb entries at and around index l in

the vector R . The amount of disturbance will be proportional to the dissimilarity between Y and the members of that semantic group.

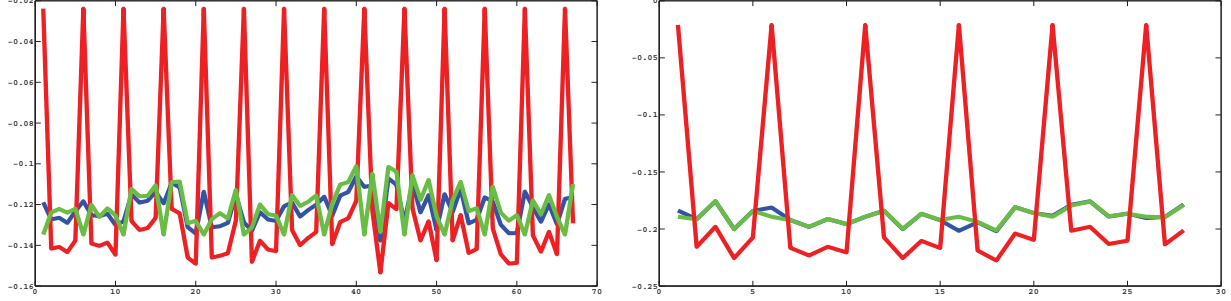


Figure 3.6: Comparison of rank-1 Tucker decomposition with visually similar and dissimilar image inserted into a tensor; Blue curve: Original Tucker decomposition vector R , Green curve: New Tucker decomposition vector R^Y with image Y inserted into the *semantic tensor* \mathbb{T}_c such that Y is visually similar to the images already contained in \mathbb{T}_c , Red curve: New decomposition vector R^Y with image Y inserted into \mathbb{T}_c such that Y is visually dissimilar to the images of \mathbb{T}_c .

To estimate the association of a test image Y with a semantic group, it is inserted at locations separated by a fixed interval, say L , in the corresponding *semantic tensor* \mathbb{T}_c by swapping images at those locations for Y . New vector R^Y is computed through Tucker decomposition. The difference between R and R^Y is an inverse measure of the association of Y with the semantic topic or *theme* represented by the semantic group T_c corresponding to tensor \mathbb{T}_c . We estimate conditional probability distribution for Y given every possible semantic group as

$$P(\mathbb{T}_c|Y) = \frac{\exp(-(R^Y - R)^T \Lambda^{-1} (R^Y - R))}{\sqrt{2\pi|I|}} \quad (3.8)$$

Λ is covariance matrix, assumed to be of form $\kappa \mathbf{I}$ where \mathbf{I} is the identity matrix and κ can be selected empirically over some held-out portion of data. This probability distribution encodes the association of the test image Y with available semantic *themes*, in turn encoding its association with the sets of *distinctive* words of each semantic *theme*.

As mentioned earlier, words occurring too frequently are given less weight in the process of forming semantic groups. To service such words, we also form a ‘general’ semantic group consisting of all training images. Each test image Y is assigned the same conditional probability, given ‘general’ semantic group, and $P(T_c|Y)$ is renormalized so that it sums to 1. Let α denote the renormalization weight which is empirically estimated by cross-validation over a held-out portion of the training dataset.

Note that no visual features have been employed in this three-step process. Instead, a comprehensive estimate of the semantic relations in terms of a probability distribution is obtained using the textual labels of the training dataset and processing of raw images through Tucker decomposition.

3.2.1.2 *Relevance Model based Image Annotation*

We devised a framework to estimate semantic *themes* or topics in terms of word groups and image tensors, and to determine the relations between the test images and these *themes*. Our argument is that such relational information contains invaluable prior knowledge regarding the word annotations for the test images. The semantic relational information regarding a test image Y is encoded in the probability distribution $P(\mathcal{T}|Y)$. Hence, this distribution needs to be incorporated in the annotation prediction process. We employed a similar relevance model based annotation prediction framework as used in the scene-based annotation model described in Section 3.1. The image is assumed to be made up of A number of visual units, i.e., $\mathbf{r} = \{r_1, r_2, \dots, r_A\}$ and the description of training image X is denoted by the set $\mathbf{w}_X = \{w_{x1}, w_{x2}, \dots, w_{xB}\}$ such that each $w_{xb} \in \mathcal{W}$ where \mathcal{W} is the vocabulary set. The size of the set \mathbf{w}_X , say B , is assumed to be the same for all the test images. The set of training images is denoted by the set \mathcal{X} of size M . Training data is divided into non-overlapping semantic groups such that each group corresponds to a *semantic tensor* \mathbb{T}_c and defines a semantic topic or *theme*.

Notice that the process of extracting semantic relational information for images is feature-independent, i.e., it employs no visual features but raw images. However, the annotation model requires the definition of each visual unit r_a . We employed grid-based visual features described in Section 3.1.2.2. The following is the generative process involved in the estimation of joint probability distribution of words and these visual units of the test image I .

1. pick a *semantic group* $\mathbb{T}_c \in \mathcal{T}$ with probability conditioned over the test image Y , i.e., $P(\mathbb{T}_c|Y)$
2. pick image X from the training set \mathcal{X} with probability $P(X|\mathbb{T}_c)$
3. for $a = 1, 2, \dots, A$
 - (a) pick a visual unit r_a from conditional probability $P_{\mathcal{R}}(.|X)$
4. for $b = 1, 2, \dots, B$
 - (a) pick a word w_b from conditional probability $P_{W_{\mathbb{T}_c}}(.|X)$

The goal of the system is to maximize the joint probability of \mathbf{r} and \mathbf{w} conditioned over Y , given by the following equation.

$$P(\mathbf{w}, \mathbf{r}|Y) = \sum_{\mathbb{T}_c \in \mathcal{T}} P(\mathbb{T}_c|Y) \sum_{X \in \mathcal{X}} P(X|\mathbb{T}_c) \prod_{b \in B} P_{W_{\mathbb{T}_c}}(w_b|X) \prod_{a \in A} P_{\mathcal{R}}(r_a|X) \quad (3.9)$$

Similar to the Equation 3.4, w_b^{th} component of the multinomial distribution of the description of the training image X is given as

$$P_{W_{\mathbb{T}_c}}(w_b|X) = \frac{\mu \delta_{w_b} + M_{w_b c}}{\mu + M_c} \quad (3.10)$$

$M_{w_b c}$ denotes the number of members of the *semantic group* T_c with word w_b in their descriptions. M_{T_c} is the total number of members of T_c . δ_{w_b} is set to 1 if the description of image X has the word w_b in it. Otherwise, It is set to 0. μ is an empirically selected constant.

Section 3.2.1.1.3 explains the estimation of $P(\mathbb{T}_c|X)$ by Equation 3.8 while $P(X|\mathbb{T}_c)$ is estimated as the following step function.

$$P(X|\mathbb{T}_c) = \begin{cases} 1/M_{T_c}, & \text{if } X \in \mathbb{T}_c \\ 0, & \text{otherwise} \end{cases} \quad (3.11)$$

$P_{\mathcal{R}}(r_a|X)$ is the density estimate for generating visual unit r_a given a training image X . Gaussian kernel is employed for this density estimate. If the training image X is assumed to be made up of a set of visual units $\{r_{x1}, r_{x2}, \dots, r_{xA}\}$, then

$$P_{\mathcal{R}}(r_a|X) = \frac{\exp(-(r_a - r_{xa})^T \Sigma^{-1} (r_a - r_{xa}))}{\sqrt{2\pi|\Sigma|}} \quad (3.12)$$

This equation uses Gaussian density kernel with covariance matrix Σ which can be taken as $\beta\mathbf{I}$ for convenience where \mathbf{I} is the identity matrix. β determines the smoothness around point r_{xa} and can be empirically selected on held-out set of the training data. This estimate signifies the importance of spatial coherence between X and Y as it compares the visual units at the same grid location, indicated by subscript a .

3.2.2 Evaluation

The evaluation metrics are mean precision and recall per word, as well as the number of words with positive recall. These evaluation measures have been popularly used for performance comparison

in various previously published papers. We employed two datasets, i.e., IAPR TC-12 and ESP, for evaluating the performance of our system against a wide variety of annotation systems. The details of these datasets are given in Section 3.1.2.1. As explained earlier, the semantic relation estimation process is feature-independent but the relevance model based annotation framework requires the description of local visual features. We tested our system with two types of visual features, i.e., the grid-based visual features, and the visual features presented by Guillaumin et al.[39]. Our system is denoted as **Tucker-AIA** and **Tucker-AIA-B** when these two types of visual features are employed, respectively. These visual features have been explained in Section 3.1.2.2.

3.2.2.1 Results

Tables 3.4 and 3.5 show the performance comparison of our system against many previously proposed strategies over IAPR TC-12 and ESP datasets, respectively. As explained in Section 2.1 of Chapter 2, different annotation strategies have their own pros and cons. CRM[64] and MBRM[30] refer to the two relevance model based systems that are computationally efficient and perform moderately well. Our systems is also based on relevance models but incorporates the semantic relational information estimated through our novel feature-independent strategy. Our systems performs much better than other relevance model based systems. TagProp[39], FastTag[18] and 2PKNN-ML[120] refer to a few iterative optimization or nearest-neighbor type frameworks. They are computationally more expansive but also produce more accurate annotations than the relevance model based systems. Our system outperforms these systems in terms of precision. The performance of our system is comparable to FastTag and TagProp-ML in terms of recall. The bulk of the computational complexity lies in the pre-processing stage of our system which involves the estimation of the semantic relations. The rest of our system is computationally efficient. Our system also beats greedy algorithms based systems such as JEC and Lasso[77]. Table 3.6 presents samples of words with very high and very low recall for both datasets.

Table 3.4: Performance evaluation for IAPR TC-12 dataset

Model	Mean precision per word	Mean recall per word	N ⁺
CRM[64]	21	15	214
MBRM[30]	21	14	186
MBRM-G[39]	24	23	223
BS-CRM[84]	22	24	250
JEC[77]	25	16	196
Lasso[77]	26	16	199
HGDM [69]	29	18	–
AP[99]	28	26	–
TagProp-ML[39]	48	25	227
TagProp[39]	46	35	266
FastTag[18]	47	26	280
2PKNN-ML[120]	54	37	278
Tucker-AIA	56	24	224
Tucker-AIA-B	61	24	242

Table 3.5: Performance evaluation for ESP-game dataset

Model	Mean precision per word	Mean recall per word	N ⁺
CRM[64]	29	19	227
MBRM[30]	21	17	218
MBRM-G[39]	18	19	209
JEC[77]	23	19	227
Lasso[77]	22	18	225
AP[99]	24	24	–
TagProp-ML[39]	49	20	213
TagProp[39]	39	27	239
FastTag[18]	46	22	247
2PKNN-ML[120]	53	27	252
Tucker-AIA	55	21	226
Tucker-AIA-B	61	20	234

Table 3.6: Sample of words with low and high recall values

IAPR TC-12	High recall	counter, fielder, root, advertising, minibus, steel, block, junction, sky
	Low recall	hair, canoe, wood, monkey, writing, grassland, green, cape, finish, face
ESP	High recall	Haryana, Europe, visa, Punjab, station, university, vegetable, Mars
	Low recall	Swing, surf, wood, crystal, cartoon, Chinese, stick, airplane, bark

3.2.2.2 Implications of Tensor Decomposition

The idea of tensor formation and decomposition has been widely explored in text mining and video analysis communities. Tensor provides a comprehensive representation for videos such that each frame of the video is a ‘slice’ in a tensor. Two out of three dimensions are representative of frame width and height while the third dimension represents time. Thus, tensors are highly suited for temporal analysis of videos. Our contribution in this work is to come up with a comprehensive tensor formation strategy for images which have no temporal connection to each other. In our case, the third dimension is used for image indices only.

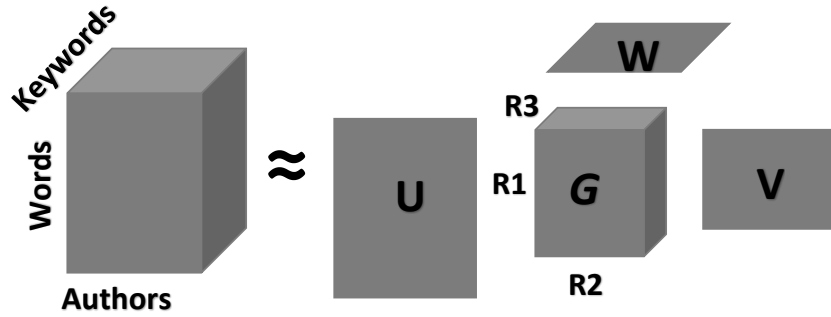


Figure 3.7: Tucker decomposition: $\mathbf{U} = \text{words} \times \text{word-groups}$, $\mathbf{V} = \text{authors} \times \text{author-groups}$, $\mathbf{W} = \text{keywords} \times \text{keyword-groups}$, $R1$, $R2$ and $R3$ represent word, author and keyword groups

Tucker decomposition of three-way tensors is a higher-order extension of Principal Component Analysis (PCA) of matrices[58]. It is a rank based estimation which results in the decomposition of the tensor in three matrices and one core tensor. The size of the core tensor is pre-specified through the rank of decomposition. Assume that the tensor is made up of documents whose authorship information and the keywords are available (as shown in Figure 3.7). In this case, the three dimensions of the given tensor represent words, authors and keywords. Three matrices are formed by decomposing this tensor, i.e., \mathbf{U} , \mathbf{V} and \mathbf{W} . The entries of matrix \mathbf{U} represent the association

of each word with every word-groups. The entries of matrix \mathbf{V} represent the association between authors and author-groups. Similarly, the entries of matrix \mathbf{W} represent the association between the keywords and keyword-groups. The association between all types of groups, i.e., word-groups, author-groups and keyword-groups, are encoded in the entries of the core tensor \mathbb{G} in Figure 3.7.

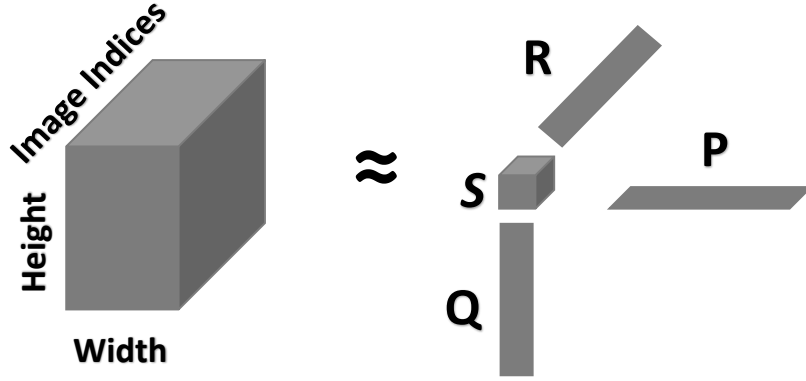


Figure 3.8: Rank-1 Tucker decomposition: S is a scalar, P , Q and R are vectors, $R = Image-indices \times 1$ where 1 represent the single *context* group represented by tensor.

Our semantic relation estimation strategy employs rank-1 Tucker decomposition. This implies that the estimated core tensor is a scalar and the estimated matrices are vectors (as shown in Figure 3.8). The idea is that the system already knows that all images in one tensor belong to one group. This is due to the fact the our system forms a tensor out of images belonging to one semantic *theme*. All of these images have similar *distinctive* words in their description, and hence are put into one group. This type of group information is potentially useful in the final task of our system, i.e., the prediction of suitable word annotations for images. The purpose of Tucker decomposition is to find out how individual elements of one group relate to the overall group so that the system may determine if some entity belongs to the group or not. Ideally, there should be little variation in the vector along the dimension of indices of images as all images are similar to each other. If a foreign entity is plugged in, this vector is perturbed. The amount of perturbation provides an estimate

of how much similar/dissimilar the foreign entity is, to the group. If the foreign entity is the test image, as in Section 3.2.1.1.3, this process estimates how much the test image is similar/dissimilar to the semantic group at hand.

3.2.2.3 Computational Complexity

The computational complexity of our semantic relations estimation scheme depends on the strategy used for Tucker decomposition. Popular existing algorithms for Tucker decomposition, such as *higher order orthogonal iterations* (HOOI)[22], are based on *alternating least square* (ALS). Phan et al. proposed a method which is computationally less expensive than HOOI[93]. ALS method is not guaranteed to converge to a global optimum or a stationary point, but if it converges under certain conditions, then it has local linear convergence rate[15]. Alternatively, differential-geometric Newton method provides convergence guarantee with quadratic local convergence rate and per iteration cost of $\mathcal{O}(H^3 D^3)$ for a tensor $\mathbb{T} \in \mathbb{R}^{H \times H \times H}$ and core tensor $\mathbb{S} \in \mathbb{R}^{D \times D \times D}$ [48].

Thorough experimentation over Tucker decomposition based semantic relational information of images indicated that tensor analysis of images is a viable tool for understanding semantic similarity between images in their raw form. Such similarity estimation alleviates the need for crafting or selection of visual features for the given collection of images. The visual features best suited to different image collections and different tasks may vary. Color and texture properties of images are important while annotating images with words but may not have significant influence over facial recognition. In addition to successfully estimating useful semantic properties of images for the task of image annotation, Tucker decomposition based analysis can find meaningful visual similarity for a wide range of image collection. We devised a Tucker decomposition based clustering scheme for raw images[113]. This clustering scheme was tested over a wide range of image collections, and it outperformed previously proposed non-parametric clustering schemes.

3.3 Multi-Layer Sparse Coding Framework for Image Annotation

Automatic image annotation is essentially a multi-label classification problem with very large number of class labels. Each word in the vocabulary set is one class label. The system has to pick a small subset of these class labels to associate with each image. The performance of the system is evaluated in term of precision and recall per word/label. The following are the formulae for precision and recall.

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \quad (3.13)$$

where TP =*true positive*, FP =*false positive*, and FN =*false negative*. If there are N number of class labels available and each image is assigned with B of these labels randomly, the probability of assigning l_{th} label to the image is $\frac{B}{N}$ while the probability of not assigning it is $\frac{N-B}{N}$. Thus the probabilities of *true positive* (TP) and *true negative* (TN) are

$$p(TP) = \frac{BM_l}{N}, \quad p(FP) = \frac{B(1 - M_l)}{N}, \quad (3.14)$$

where M_l is the fraction of images whose ground truth labels include the l_{th} label. Similarly, the probability of *false negative* (FN) is

$$p(FN) = \frac{M_l(N - B)}{N}, \quad (3.15)$$

Plugging in these probability estimates into the formulae of precision and recall gives us the following

$$Precision \propto M_l, \quad Recall \propto \frac{B}{N}, \quad (3.16)$$

In general, $B \ll N$. Thus,

$$Precision \propto M_l, \quad \text{whereas} \quad Recall \propto \frac{1}{N} \quad (3.17)$$

Hence, the precision for any label is directly proportional to its frequency and recall is inversely proportional to the total number of available labels, for a model that assigns classification labels randomly. It implies that the annotation model which is a classification model by its nature, suffers from low recall because of the large number of available classification labels. This observation is validated by the results reported in Sections 3.1.2.3 and 3.2.2.1. Not only our relevance model based methods but also a large variety of previously proposed systems report much lower recall scores than the precision scores.

The precision and recall analysis that we described above also implies that the systems are inclined towards being highly precise for very frequent labels. The overall performance is reported as average precision over all labels. Every label is given equal importance in such an evaluation scenario. Systems can justifiably report high precision scores by being highly precise for frequent labels only. The relation between the information content and the frequency of words has been studied in the field of text mining, especially in terms of the text retrieval problem. In terms of information theory, moderately frequent words contain more information content than extremely frequent words. Such words represent '*surprising*' events of high information content, and are very important in search queries for text retrieval engines[117]. Hence, a system that reports high precision scores by focusing on extremely frequent words only, ignore words of high information content which are very valuable for search and retrieval systems. Since image retrieval engines are the intended application tools for image annotation systems, ignoring the words with high information content is problematic.

We devised a model that attempts at being precise for words with a wide range of frequency, as

well as achieving high recall in addition to high precision. Our system tackles the problem of low recall at its root, i.e., large number of available labels, by establishing semantic contextual relations for images. The system identifies a set of *themes* or semantic topics in the training data using an approach similar to the one employed in Section 3.2. Later, two layer of sparse coding are employed to establish relations between images and the semantic topics, and to predict individual labels or tags for images. The sparse coding layer that actually predicts these tags deals with a smaller set of all available labels, intelligently reduced by the first layer of sparse coding in reference to the semantic relations of the given image.

Sparse coding is the process of learning a sparse representation of a signal in terms of coefficients of a set of basis signals or predictor variables. Tibshirani proposed LASSO (least absolute shrinkage and selection operator) that includes an ℓ_1 -norm penalty to induce sparsity and interpretability in the learned model [116]. Variations of this model have been proposed to incorporate any inherent structure among the predictor variables [105]. Such modeling is beneficial when the group structure carries semantic meaning.

The idea of sparse modeling has been explored widely in computer vision and image processing communities for selecting visual features that can strongly predict the target label. In such cases, visual features are the predictor variables and the sparse coding model assigns high weights only to the features with strong correlation to the final label of the task at hand. The sparsity constraint in such modeling schemes ensures that the highly relevant visual features are identified while the rest are assigned zero weights[71, 13, 37, 47, 19, 40, 76]. Image classification and face recognition systems have been built using sparse coding models that treat individual images, rather than visual features, as predictor variables[123, 36]. The number of class labels for such systems is usually much smaller than a typical labels' set for an annotation system. Systems dealing with image classification concurrently with image annotation are limited in their application as they require training data to have class labels, in addition to their word annotations. Class labels set is usually

much smaller than the set of all possible annotations[36]. The system described in [36] employs such class labels to induce group structure among predictor variables, i.e., the training images. On the other hand, our system deals with large labels' set problem of image annotation, treats individual training images as predictor variables and induces group structure among predictor variables without requiring any additional input. We named our system 'MultiSC-AIA'.

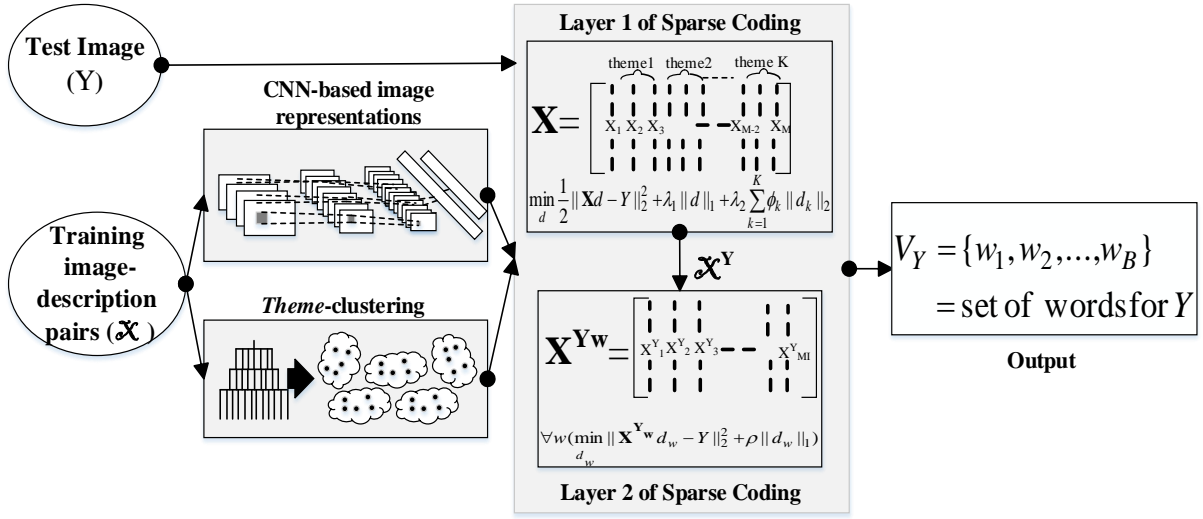


Figure 3.9: System Architecture; \mathcal{X} denotes the set of training items. \mathcal{X}^Y represents the subset of training items that belong to *themes* selected for test image Y . V_Y is the set of words selected for test image Y .

3.3.1 System Architecture

Figure 3.9 provides an overview of our system. As described at the start of this chapter, set \mathcal{X} of size M contains training data in the form of image-description pairs. Vocabulary set \mathcal{W} of size N contains all available labels for the problem. The system divides the training set \mathcal{X} into non-overlapping groups such that the group \mathcal{X}_k corresponds to one semantic *theme* C_k . \mathcal{C} denotes the set of all *themes*. When provided with a test image Y , the first layer of the sparse coding identifies

the *themes* related to the image Y . Let \mathcal{C}^Y (where $\mathcal{C}^Y \subset \mathcal{C}$) denote the set of the related *themes*. The second layer of the sparse coding is fed with a subset of training images $\mathcal{X}^Y \subset \mathcal{X}$. This subset consists of all training images belonging to the *themes* from the set \mathcal{C}^Y ($\mathcal{X}^Y = \{X | X \in \mathcal{X}_k \wedge C_k \in \mathcal{C}^Y\}$). The set \mathcal{W}^Y contains all the words used in the descriptions of $X \in \mathcal{X}^Y$, and thus $\mathcal{W}^Y \subset \mathcal{W}$. This is an intelligently reduced labels or vocabulary set which helps our system achieve high recall values while still being highly precise. The second layer of the system builds a set V_Y such that $V_Y \subset \mathcal{W}^Y \subset \mathcal{W}$ and V_Y contains appropriate tags for image Y .

3.3.1.1 Visual Feature Extraction

Deep convolutional neural network, inspired by the network proposed by LeCun et al.[65], have gained immense popularity for representation learning from raw images. ImageNet is a vast collection of images where each image is a member of some class based on the object shown in the image [23]. This large database of labeled images provides an excellent training dataset for deep convolutional neural networks (CNN). We employed the CNN proposed by Krizhevsky et al.[59], and trained over ImageNet, for learning image representations. The final output of this network corresponds to 1000 ImageNet labels. Each image is represented by 4096-dimensional coefficients vector obtained from the last fully connected (fc7) layer.

We argue that the image representations learned by ImageNet-trained deep CNN complement the *theme*-structure of the dataset. As explained in the following section, the *themes* in the training dataset are determined by the words associated with the training images. In standard image annotation datasets, names of the objects present in the images constitute a large portion of these words. Thus, the tags are of similar nature as the class labels of ImageNet database. As a result, the images of one *theme* share substantial similarity in their visual features extracted from ImageNet-trained CNN.

3.3.1.2 Theme-based Clustering



Figure 3.10: A sample *theme* from Flickr30K dataset; ‘ person playing banjo’ seems to be the distinctive characteristic of this *theme*.

Our system intelligently reduces the size of the labels set/vocabulary set with respect to the *themes* of the given test image without needing any additional input. The *themes* are extracted from the training dataset by the same process as employed by our Tucker-AIA model described in Section 3.2. Images described by the same set of *distinctive* words are expected to describe the same *theme* in their visual contents. The *tfIdf* representation of image descriptions promises to assign higher weights to the *distinctive* words rather than the highly frequent words. Training images are clustered through hierarchical clustering scheme with respect to the cosine similarity between the *tfIdf* vectors of their textual descriptions. Each cluster contains images that share the some *distinctive* words in their descriptions as well as a common *theme* in their visual contents. Sample *themes* for IAPR TC-12 datasets are shown in Figures 3.3 and 3.4. Figure 3.10 shows a sample *theme* for Flickr30K dataset. These *themes* carry rich semantic and contextual meaning. As explained in Section 3.3.1.1, images of the same *theme* also share similarity in their CNN features.

3.3.1.3 Multi-Layer Sparse Coding

The system employs two layers of sparse coding to predict appropriate words/tags for images. The first layer identifies *themes* relevant to the test image Y and the next layer predicts the annotations

for Y in the light of its relevant semantic *themes*. The following sections describe the formation of each layer one-by-one.

3.3.1.3.1 Group sparse coding for theme identification

The first layer of sparse coding is inspired by a variation of lasso modeling that incorporates an inherent group structure among the predictor variables[105]. In our approach, the training images are treated as predictor variables and the *themes* of the training images define the group structure of the predictor variables. The goal of this layer is to identify a set of appropriate *themes* for the test image. As explained in Section 3.3.1.2, *themes* defined over the training data carry rich semantic and contextual meaning. It is intuitive that a semantic background for the test image is defined when *themes* of this test image are identified.

Training images in the form of their CNN-learned representations are used as the set of basis vectors. The data matrix, denoted by \mathbf{X} , is built such that each column contains one training image. The adjacent columns are grouped such that images contained in one group of columns belong to the same *theme*. Y denotes the CNN-learned representation of the test image. The goal is to minimize the following cost function.

$$\min_{\mathbf{d}} (||\mathbf{X}\mathbf{d} - Y||_2^2 + \lambda_1 ||\mathbf{d}||_1 + \lambda_2 \sum_{k=1}^K (\phi_k ||\mathbf{d}_k||_2)) \quad (3.18)$$

This objective function is regularized by two penalty functions. A penalty based on ℓ_1 -norm of the coefficients vector \mathbf{d} induces sparsity in the learned model. Sparsity at the group-level is induced by a penalty of ℓ_2 -norm of the groups of coefficients of every *theme*. This penalty ensures that the members of the same group or the images of one *theme* are weighted in accordance with all members of the group and as few groups/*themes* are assigned non-zero coefficients as possible.

Weights assigned to the two penalty terms, i.e., λ_1 and λ_2 , indicate the emphasis put on each penalty term by the framework. In our approach, equal emphasis is put on all groups/*themes*, i.e., $\forall k, \phi_k = 1$




















Appropriate *themes* for the test image Y are the ones whose corresponding groups of coefficients are assigned non-zero values. The set of these *themes* is denoted by \mathcal{C}^Y . A subset of training images (\mathcal{X}^Y) is prepared such that m_{th} training image belong to \mathcal{X}^Y if it belongs to one of the *themes* of the set \mathcal{C}^Y and $d_m > 0$ for the optimal d . The words used in the descriptions of images of the set \mathcal{X}^Y for the set $\mathcal{W}^Y \subset \mathcal{W}$. The subset \mathcal{W}^Y is the intelligently reduced labels set, specific to the *themes* of the test image. When annotations are being predicted at the next layer of sparse coding framework, adverse effects of the large size of the labels' set on the recall of the system are reduced as this smaller subset is being considered as the labels' set.

To prevent the precision from dropping, it is necessary that the subset \mathcal{W}^Y contains all the labels or the words that can be potentially related to the image Y . Association of the image Y with multiple *themes* avoids unnecessary limiting of the set \mathcal{W}^Y . The test image may show characteristics of multiple *themes* and should be processed in the light of all these *themes*. Each row of the Table 3.7 shows one test image along with a few sample images from its relevant *themes*. For example, the test image in the third row shows a guitarist, a guitar and a drum set. It has one relevant *theme* showing guitarists with their guitars and one showing drummers and drum sets. Ignoring any of these semantic *themes* would result in a *theme*-dependent vocabulary set \mathcal{W}^Y that does not contain all potential tags for the test image Y .

3.3.1.3.2 Regularized linear regression modeling for tag prediction

After the identification of appropriate *themes* and the resulting subset of the training set \mathcal{X}^Y , a second sparse coding layer is used to find the set of appropriate annotation V_Y for the image Y .

Table 3.7: Test images and training images from multiple *themes* related to them; samples from different *themes* are separated by bold vertical lines.

Test Image	Sample images from <i>themes</i> related to the test image											
												
												
												
												

Each word $w \in \mathcal{W}^Y$ has a set of representative images $\mathcal{X}^{Y_w} \subset \mathcal{X}^Y$ such that word w occurs in the description of each image of this representative set. The matrix \mathbf{X}^{Y_w} is built such that each column of this matrix correspond to one image of the set \mathcal{X}^{Y_w} . The following objective function is minimized for every word $w \in \mathcal{W}^Y$.

$$\min_{\mathbf{d}_w} (||\mathbf{X}^{Y_w} \mathbf{d}_w - Y||_2^2 + \rho ||\mathbf{d}_w||_1) \quad (3.19)$$

This objective function represents a linear regression model with ℓ_1 -norm of coefficients vector \mathbf{d}_w as penalty that aims at inducing sparsity in the learned model.

The target variable is the test image Y itself. Therefore, the task of the learned model is the reconstruction of the test image Y using the weighted linear combination of the predictor variables, i.e., the member images of the set \mathcal{X}^{Y_w} . The quality of the reconstruction is judged by the cosine

similarity as

$$sim^{Y_w} = \frac{(\mathbf{X}^{\mathbf{Y}_w} \mathbf{d}_w) \cdot \mathbf{Y}}{\|\mathbf{X}^{\mathbf{Y}_w} \mathbf{d}_w\| \times \|\mathbf{Y}\|} \quad (3.20)$$

Higher value of sim^{Y_w} indicates that the representative images of the word w can reconstruct the test image Y with low error. It implies that the image Y is visually similar to the training images that are tagged with the word w . Therefore, the image Y should also be annotated with the word w . One such model is learned for every word $w \in \mathcal{W}^Y$. If V_Y is the set of B annotations for the image Y , then V_Y contains the first B words if the words are sorted in descending order of their sim^{Y_w} scores⁴.

3.3.2 Evaluation

Mean precision per word, mean recall per word and the number of words with positive recall are used as evaluation measures. Since our system is aimed at maintaining a balance between precision and recall values, F-score is also considered as an evaluation measure. F-score is the harmonic mean of precision and recall and indicates the trade-off between these two measures. F-score has been previously used for evaluation of various annotation system⁵[120].

In addition to IAPR TC-12 and ESP game datasets that have been used for evaluation of Scene-AIA and Tucker-AIA (Sections 3.1 and 3.2, respectively), Flickr30K dataset was employed for evaluation of this system. The Flickr30K⁶ [127] dataset contains approximately 31,000 images collected from Flickr⁷. These images show people engaged in everyday activities. Each image is

⁴We employed the SLEP software package <http://www.yelab.net/software/SLEP/> for the training of the sparse coding models at both layers of the system.

⁵Mean F-measure has been defined as the harmonic mean of the mean precision and the mean recall, i.e., $F = \frac{2PR}{P+R}$ [120]. We use the same definition.

⁶<http://shannon.cs.illinois.edu/DenotationGraph/>

⁷www.flickr.com

associated with 5 captions collected through crowd-sourcing. This dataset has been previously used to test description retrieval systems and the systems generating sentence-like captions[54, 121, 79]. We annotated this dataset with individual words through MultiSC-AIA as well as some previously proposed annotation methods, to evaluate how different methods adapt to the datasets with different characteristics. Testing is performed over 1000 randomly picked images from the dataset. We used TreeTagger⁸ to tokenize, lemmatize, and part-of-speech tag image captions. Since every image is associated with 5 captions, every word that is present in more than one caption is taken as a valid annotation for the image. The vocabulary set \mathcal{W} consists of 316 frequently occurring nouns, verbs and adjectives.

First, hierarchical clustering is performed to divide the training images into clusters/*themes* for all three datasets as described in Section 3.3.1.2. This clustering is based on the cosine similarity between *tfl**df* representations of the textual descriptions of images over the extended vocabulary sets. Clusters with too few members are dropped such that the remaining clusters contain approximately 90% of the dataset. This process results in 771, 1346 and 1102 *themes* for IAPR TC-12, ESP game and Flickr30K datasets, respectively.

3.3.2.1 Results

Tables 3.8, 3.9 and 3.10 show the performance of various image annotation systems over IAPR TC-12, ESP game and Flickr30K datasets, respectively. Rows of tables 3.8 and 3.9 are grouped in terms of the approach of the methods. The first group of rows contain methods based on the relevance model from the domain of machine translation. The second group contains methods using the nearest-neighbor type algorithms. The third group contains systems using a variety of approaches such as greedy label transfer[77], random forests[35] and deep neural networks[57].

⁸<http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger>

The fourth group shows the performance of some previously proposed systems with fc7 image features, i.e., the image representations extracted from the last fully connected layer of ImageNet-trained CNN described in citeAlexNet. MultiSC-AIA is also part of this group as it also uses fc7 image features. Flickr30K dataset was introduced relatively recently as compared to IAPR TC-12 and ESP game datasets. Therefore, various image annotation papers have not reported results over this dataset. We evaluated various previously proposed systems using different approaches in addition to our MultiSC-AIA model over this dataset.

Table 3.8: Performance evaluation for IAPR-TC-12 dataset

Model	Mean precision per word	Mean recall per word	Mean F-score	N⁺
CRM[64]	21	15	18	214
MBRM[30]	21	14	17	186
BS-CRM[84]	22	24	23	—
Scene-AIA	56	25	35	230
TagProp-ML[39]	48	25	33	227
TagProp[39]	46	35	40	266
FastTag[18]	47	26	34	280
2PKNN-ML[120]	54	37	44	278
JEC[77]	25	16	20	196
Lasso[77]	26	16	20	199
HGDM [69]	29	18	22	—
AP[99]	28	26	27	—
Deep rep.[57]	42	29	34	252
Random Forests[35]	45	31	37	253
Scene-AIA(fc7)	63	27	38	259
TagProp(fc7)	32	40	36	264
MultiSC-AIA(fc7)	41	42	42	250

It is important to note that the precision can be increased for many systems at the cost of the decrease in recall and vice versa. On the other hand, F-measure is the harmonic mean of the precision and the recall, providing a unified evaluation measure that incorporates the trade-off between the precision and the recall. As shown in Tables 3.8, 3.9 and 3.10, our method outperforms all previously proposed approaches in terms of mean recall per word for all datasets. Our method

also outperforms all other methods in terms of mean F-measure for ESP game and Flickr30K dataset. It performs comparable to the best performing previously proposed approach (2PKNN-ML[120]) for the IAPR TC-12 dataset, while outperforming all other methods in terms of mean F-measure. The precision and the recall values achieved by our system, are very close to each other for all three datasets, indicating that our method maintains its designed characteristics for the datasets with different characteristics.

Table 3.9: Performance evaluation for ESP game dataset

Model	Mean precision per word	Mean recall per word	Mean F-score	N ⁺
CRM[64]	29	19	18	227
MBRM[30]	21	17	19	218
Scene-AIA	60	20	30	234
TagProp-ML[39]	49	20	28	213
TagProp[39]	39	27	32	239
FastTag[18]	46	22	30	247
2PKNN-ML[120]	53	27	36	253
JEC[77]	23	19	21	227
Lasso[77]	22	18	21	225
AP[99]	24	24	24	—
Deep rep.[57]	38	22	28	228
Random Forests[35]	45	24	31	239
Scene-AIA(fc7)	61	21	31	245
TagProp(fc7)	20	37	33	245
MultiSC-AIA(fc7)	39	38	39	237

A variety of visual features were evaluated previously by image annotation systems, ranging from grid-based visual features[30] to normalized-cuts based *blobs*[49]. Our method includes a deep convolutional neural network at the initial processing stage to automatically learn visual representations for images. Such image representation has greatly improved the performance of image classification and character recognition systems[59, 106, 65]. Our approach reaps the rewards of effective image representation learning by using deep convolutional neural networks. To thoroughly assess the benefits of our multi-layer sparse coding scheme, above and beyond the benefits

of the chosen image representation scheme, we tested the previously proposed methods such as Scene-AIA[111] and TagProp[39] with the same image representation (denoted by ‘fc7’) as our classifier employs for all three datasets. It is obvious from Tables 3.8, 3.9 and 3.10 that our classifier outperforms previous methods even when they use the same image representation. This fact points to the effectiveness of our sparse coding framework. TagProp is based on a nearest-neighbor algorithm and was originally used with images represented by a combination of local and holistic features in [39]. For ‘fc7’ features, this system achieves higher recall and comparatively low precision as compared to its application over visual features described in [39] for IAPR TC-12 and ESP game datasets. For both types of image representations, precision and recall values are vastly different from each other and the F-measure is less than the value achieved by our classifier.

Table 3.10: Performance evaluation for Flickr30K dataset

Model	Mean precision per word	Mean recall per word	Mean F-score	N ⁺
MBRM	16	23	19	264
Scene-AIA(fc7)	35	18	24	179
TagProp(fc7)	23	28	25	243
MultiSC-AIA(fc7)	26	27	27	243

Table 3.11: Sample high and low recall words from three all datasets; The words with high frequency (e.g., ‘man’, ‘woman’, ‘dog’) or related to distinctive visual *themes* (e.g., *tennis match* or *bicycle race*) achieve better recall.

Dataset	Word with high recall	Words with low recall
IAPR TC-12	court, net, player, tennis, sky, cyclist, spectator	backpack, pavement, green, clothes, hedge, garden, stand
ESP	man, white, airplane, coin, red, sky, black, tree	mouth, CD, child, swim, statue, hill, shadow, school, floor
Flickr30K	wave, man, woman, soccer, dog, shirt, people	paint, striped, sunglass, arm, dark, vest, watching, long

Table 3.11 shows a few samples words assigned very high and very low recall values by our system for all three datasets. It is evident that, in addition to frequently occurring words, words related

to distinct visual *themes* or scenes (such as *tennis match*, *soccer match* and *bicycle race*) achieve better recall. This property can be attributed to the *theme* selection process of our framework.

3.3.2.2 Noise Reduction














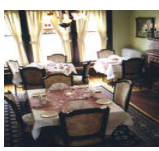

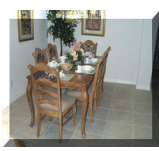











The first layer of sparse coding identifies the *themes* related to the test image Y . As explained in Section 3.3.1.2, the training images of each *theme* are not only visually similar but also share similarity in their CNN features. When our framework deems a *theme* relevant to the test image Y , it implies that the member training images of the *theme* and the test image share similarity in their visual contents as well as their CNN features. Only these training images and the words used in their descriptions (the sets \mathcal{X}^Y and \mathcal{W}^Y , respectively) are passed to the second layer of sparse coding, effectively reducing the search space for this layer of processing.

The second layer of sparse coding processes every word $w \in \mathcal{W}^Y$ separately. All images of the set \mathcal{X}^Y tagged with the word w , are used to reconstruct the test image Y . One word does not limit the variety of visual content or the overall appearance of the image by any significant amount. The same word or annotation can be associated with a wide variety of images. Only a few of these images are similar to the test image Y in terms of the contents and the visual feature vectors. The rest of the images act as noise while reconstruction model is being estimated. The first layer of our sparse coding framework acts as a filter for this noise as it passes only the images that are visually similar to the test image Y , to the second layer. Therefore, the reconstruction modeling at the second layer has to deal with less noise, resulting in quick and accurate model estimation.

Table 3.12 shows a few examples of this phenomenon. For each row, the first column contains the word w , the second column shows the test image Y , the third column displays the training images labeled with w and visually similar to Y , and the last column contains images labeled with w but visually dissimilar to the image Y . An accurate linear regression model can be fitted over images

in the third column for the test image Y . If the images from the fourth column are added to this model, they act as noise which the estimation process has to filter out. In our method, the first layer of sparse coding can potentially filter this noise as these images belong to *themes* that seem inappropriate for image Y and are not likely to be selected.

Table 3.12: Noise in training data for individual words

Word (w)	Test Image	Relevant training images with word w				Irrelevant training images of word w (Noise)		
Car								
Sea								
Table								
Airplane								

3.3.2.3 Time Complexity

Using multiple layers of sparse coding reduces the time complexity of the system. After identifying the *themes* for the test image, the second layer of sparse coding has to deal with only subsets of the training data \mathcal{X} and that of vocabulary set \mathcal{W} , denoted by \mathcal{X}^Y and \mathcal{W}^Y , respectively. These subsets consist of training images belonging to the *themes* selected for the test image and the unique words

used in their descriptions, respectively. The processing of the second layer of sparse coding is sped up because of this subset selection.

A single layer sparse coding framework is essentially a special degenerate case of our model obtained by assuming that all of the training data belongs to one *theme*. Consequently, this one *theme* is appropriate for every test image Y and all of the training images and their vocabulary are passed to the next layer of sparse coding ($\mathcal{X}^Y = \mathcal{X}$ and $\mathcal{W}^Y = \mathcal{W}$). The system would then need to learn a regularized linear regression model for every word of the vocabulary \mathcal{W} over all training images associated with that word. According to our experiments, the processing time per image for such a degenerate case can be up to an order of magnitude more than our multi-layer model, depending on the regularization parameters of the system. Annotation accuracy for such a model also decreases, in comparison to our model, when more challenging datasets such as Flickr30K, are used for testing. Flickr30K dataset is larger and contains images of wider visual variety as compared to IAPR TC-12 and ESP game datasets. This implies that the training data for each word w contains more noise (images visually dissimilar to the test image), deteriorating the accuracy of learned models.

3.3.2.4 Theme Selection and Image Organization

As explained in Section 3.3.1.3.1, a test image Y may be associated with multiple *themes*. We observed that multiple *themes* associated with the given test image Y usually represent multiple aspects of that image. No one *theme* can describe the image in totality but the combination of these *themes* successfully pinpoint the contents of the given image. The selected *themes*, in addition to the annotated words, have huge potential for image database organization and management systems. Images can be linked to each other through common *themes* for easy access. Each link will represent certain aspect of the visual contents shared between the connected images. Such links may also be beneficial for a system that needs fast access to images with some common

aspect of visual contents. New images may also be integrated with an existing database of linked images. Aspects of visual contents of the new images may be identified by our sparse coding model and used to link the new image to the existing images in the database.

Figure 3.11 shows an example of a test image and sample images from four of its relevant *themes*. This example clearly shows that multiple *themes* cover multiple aspects of the visual contents of the test image.

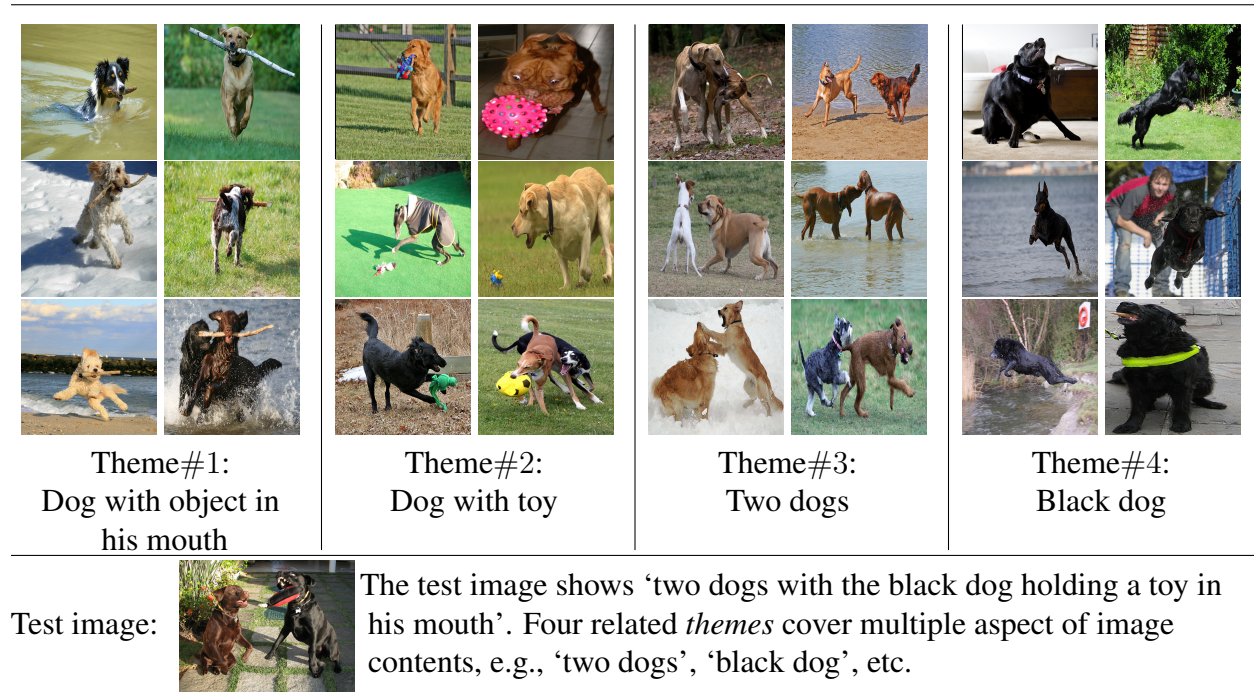


Figure 3.11: Sample images from multiple *themes* associated with the given test image

3.3.2.5 Precision for Descriptive Words

Figure 3.12 show precision-frequency plots of our MultiSC-AIA model and the previously proposed MBRM[30] model. The annotations are sorted in the ascending order of frequency for IAPR TC-12 dataset. Mean precision and mean frequency of the sets of every 10 annotations

are calculated. Curves fitted through the points representing these tuples (y-axis:precision, x-axis:frequency) are called the precision-frequency plots. Our system maintains high precision for words with a wide range of frequency while MBRM is highly precise for very frequent words only. The behavior of MBRM is in compliance with the implications of Equation 3.17.

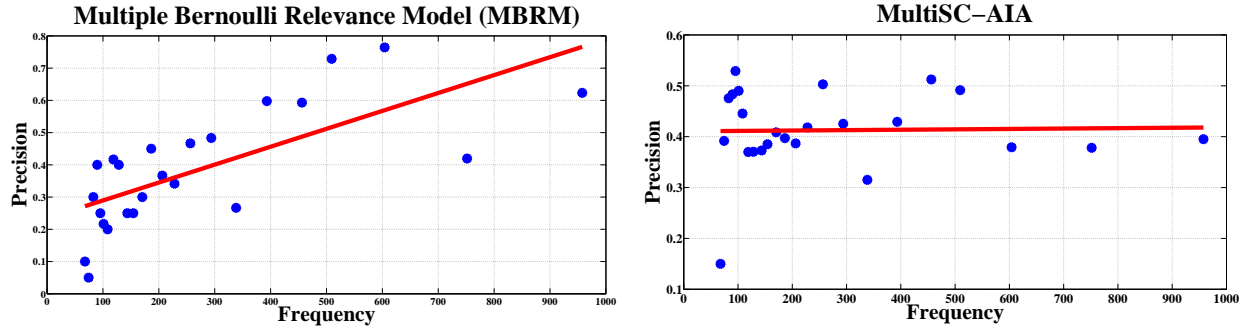


Figure 3.12: Precision vs. Frequency of MBRM and MultiSC-AIA (IAPR TC-12)

In addition to outperforming other methods in overall performance, our system is also more precise for highly descriptive words of moderate frequency (‘tennis’, ‘stadium’, ‘waterfall’, ‘cathedral’, ‘player’). Such words are very important in search and retrieval scenario as, in terms of information theory, they represent ‘*surprising*’ events with more information content than the ‘*expected*’ events depicted by highly frequent words (‘sky’, ‘man’, ‘wall’)[117]. Most of the other systems tend to rely on precision of highly frequent words to improve overall performance.

The proposed multi-layer sparse coding framework takes advantage of the visual representations learned by ImageNet-trained CNN which strongly correlates with ground truth annotation for various image annotation benchmark datasets. With the use of multiple layers of sparse coding the proposed framework overcomes an inherent problem in annotation models, i.e., low recall because of large number of labels. Thorough experimentation clearly indicates that our system maintains its signature characteristics for a variety of annotation benchmark datasets.

3.4 Conclusion

In this chapter, we explained various systems that we devised for automatic image annotation. Automatic annotation of images with descriptive text is an important problem involving multi-modality datasets. Such systems have huge potential benefits for image search and retrieval engines. The lack of correlation between the visual features and the words, i.e., *semantic gap*, is the main challenge faced by such systems. Our core strategy for bridging this gap involves understanding of semantic relations of images. Such semantic relations constitute the prior knowledge needed to annotate them with individual words.

Our strategy involves automatic extraction of semantic topics or *themes* from the training data. We implemented such semantic topic extraction based on scene analysis of images as well as co-occurrence patterns of words. Later, images are grouped such that each group becomes the visual representation of one semantic topic or *theme*. When a test image is encountered, its association with each of these semantic topics is estimated based on the similarity between the visual contents of the test image and the visual representations of semantic *themes*. We experimented with non-parametric Gaussian kernel, tensor analysis and sparse structured coding to estimate the relations between the images and semantic topics. These relations constitute the semantic relational information that we incorporate in annotation prediction models to generate meaningful annotations for images.

We explored two modeling schemes for predicting annotations for test images, in reference to their semantic relations. Our first modeling scheme involves a relevance model inspired expectation process over the training data. This expectation process is sensitive to the semantic relational information. The semantic relational information is encoded in a probability distribution over all available semantic *themes* for the test and the training images. Our second modeling scheme involves multiple layers of sparse coding framework. The first layer of sparse coding incorporates

the semantic *theme* information as the group structure of the predictor variables. The second layer predicts the annotations for the test image in light of the semantic *themes* related to the image. Thorough experimentation has shown that our idea of incorporations of semantic relational information is effective in predicting semantically meaningful annotations for images, in addition to being computationally efficient.

CHAPTER 4: CROSS-MEDIA SEMANTIC RELATIONS FOR NEWS MATERIAL

In this chapter, we introduce our ideas for machine understanding of cross-media semantic relations for news collections; especially the relations between news images and the words and linguistic features of the text.

News collections are abundantly available through the websites of print and electronic news media sources. These collections include a wide variety of data modalities. Therefore, knowledge mining in news datasets needs to push the boundaries of cross-media relation extraction. In addition to the images and their short textual descriptions (i.e., captions), long text sequences are also part of the news datasets in the form of news articles. News items are usually associated with some structured textual information in the form of news category labels and article keywords. News items are also assigned timestamps and titles. Many linguistic features, such as named entities, are implicitly available in news articles. To understand the relations between words and image for such dataset, all these information sources need to be considered in reference to each other.

We expand the scope of the core idea of this dissertation, i.e., semantic relation building for understanding image-text relations, to understand the images and the text of news datasets. We devise an automatic image description generation framework for news images. Our framework aims at automatically generating image descriptions which match actual the real world captions of news images. As explained in Section 4.2.1 in Chapter 1, real world image captions involve hints to the *context* of images, in addition to the description of visual contents of images. As for any type of images, visual contents of news images contain invaluable hints to its *context*. But contextual information for news images may also be encoded in auxiliary information sources or sources *extrinsic* to the news images. Such auxiliary source include news articles, category labels, article keywords,

etc. We explore the application of our idea of semantic relation extraction to gather contextual cues for news images from every possible source, be it *intrinsic* (e.g., semantic scene characteristics of images) or *extrinsic* (e.g., news article, metadata) to the images. We incorporate such semantic contextual cues as prior knowledge in our news image description generation system.

Propagation of semantic contextual information from heterogeneous sources requires identification of a common representation scheme. We described our image annotation models in Sections 3.1 and 3.2 of Chapter 3 that encode semantic relational information in probability distributions. We argue that the probability space is an excellent common representation space for semantic contextual information if we can devise a framework to encode such information from every source in a probability distribution. In this chapter, we introduce our ideas for estimating probability distributions representing semantic contextual information from every information source included in the news collections. We also present a flexible framework to incorporate this probabilistic semantic contextual information in the process of automatic image annotation.

We also devise a framework to generate sentence-like captions for news images that can potentially reduce the human effort required in writing descriptions for news images. We base our framework for such caption generation on the ideas of *extractive* summarization from the field of text mining[52, 78]. Since large amount of grammatically correct text is available with each image in the form of a news article, the goal of the caption generation systems is to *extract* the best text sequence to describe the image.

The final goal of understanding relations between words and images is to automate deep understanding of large, heterogeneous datasets available on the internet. In case of news datasets, vast amount of semantic information is available in news articles. It is necessary for the system to be able to develop in-depth understanding of this semantic information to fully encompass the scope of information available in news datasets. News articles contain a wide variety of words. Some

of these words have special meaning. For example, words that indicate ‘named entities’ are a very important linguistic feature. Named entities are the names of people, places and organizations. Studies have shown that such named entities are the most common query words for news and blogs databases[82]. Various computer vision problems deal with the recognition and linking of such entities with the visual contents of images. Examples of such problems include the recognition of faces in news images (persons)[5, 89], the detection of the company logos and the trademarks in images (organizations)[29, 101, 55], and the identification of landmarks in natural scenes (places)[16, 17, 1]. Therefore, developing deep understanding of these named entities and their semantic relationships is very important to expand the scope of image-text relations. News images contain some semantic information regarding these named entities, but it is essential to thoroughly process the semantic information encoded in articles to develop better understanding of such entities. We devise a system for automatic understanding of semantic relations among named entities by processing the text of news articles.

For detailed exploration of our ideas, we collected our own news dataset from the website of TIME magazine. This dataset and its characteristics are described in detail in the following section.

4.1 News Dataset

One of our contributions is the collection of a sizable dataset of news images, i.e., the TIME dataset. We collected 19841 articles from the website of the TIME magazine¹. We ensured that each downloaded article is associated with one image and the caption of the image is also available. We also collected the titles of these articles. News articles are organized into news categories. News sources also assign keywords to news articles. We collected these category labels, keywords and publication timestamps for all of 19841 image-caption-article tuples.

¹www.time.com

The text of articles and captions was tokenized, lemmatized, and part-of-speech tagged using Tree-Tagger. The vocabulary set consist of nouns, verbs and adjectives only. The vocabulary set for articles contains 6350 unique words, each with frequency more than 100. Every word with frequency more than 20 is collected to form the vocabulary set for captions. The size of this vocabulary set is 1937. Collective vocabulary set has 6449 unique words.

There are 10 unique news categories and 719 unique keywords. Figures 4.6 and 4.7 show the distribution of the data across news categories and keywords, respectively. These distributions are uneven but beyond the control of the system developers. Average lengths of articles, captions, and titles are 163, 10 and 5 words, respectively, in terms of the selected vocabulary set. Average image-size is 480×320 pixels.

Feng et al. presented a small dataset, called the BBC dataset of news images along with their captions and articles [31]. They manually evaluated the dataset to conclude that nouns, verbs and adjectives mentioned in an image caption are considered ‘relevant’ to the image as annotations by humans [33]. This implies that the news images and their captions form an excellent benchmark to test image annotation and caption generation frameworks. There is no additional effort required to collect human-written descriptions of news images, unlike the images available in datasets such as IAPR TC-12, MSCOCO, etc. Our dataset is six times larger than the BBC dataset. It contains, in addition to images and captions, a variety of auxiliary information sources (e.g., article, title, keywords, news categories). Availability of such auxiliary information sources enables us to test our hypothesis that both images and auxiliary information sources define the *context* for images. Image description generation systems need to incorporate this *context* information to improve the quality of automatically generated annotations and captions for images.

It is also worth noticing that the datasets like IAPR TC-12 and MSCOCO have artificial image descriptions in the sense that the captions available in these datasets were not actually associated

with the images in the real world. Captions of news images are the real world descriptions of the corresponding images. Hence, the news image-caption pairs provide an opportunity to test the applicability of any annotation system in the real world. We also evaluate certain aspects of our news image description generation framework on the standard image annotation datasets like IAPR TC-12 ESP, Flickr30K and MSCOCO to demonstrate the characteristic differences between these datasets and our TIME dataset. IAPR TC-12 and ESP datasets are explained in Section 3.1.2.1 of Chapter 3. Details of Flickr30K have been discussed in Section 3.3.2 in Chapter 3. MSCOCO dataset contains approximately 124000 Flickr images. Human annotators with no knowledge regarding the *context* or the background story of images were asked to write captions for images in this dataset. After tokenization and part-of-speech tagging, only frequently occurring nouns, verbs and adjectives are included in the vocabulary set for this datasets. The approximate size of the vocabulary set is 300. We employed the same split of data between test and training sets as employed by previously published papers involving standard image annotation datasets.

We use \mathcal{X} as the notation for the training subset of the TIME dataset. Each item $X \in \mathcal{X}$ consists of an image (X^I), an article (X^d), a news category label (X^n), and a keyword (X^{key}). The collective vocabulary set is denoted by \mathcal{W} . Sizes of the training set \mathcal{X} and the vocabulary set \mathcal{W} are denoted by M and N , respectively. The textual description of image X in the form of a sentence is denoted by H_X . For semantic context extraction, the dataset is divided into a set of semantic groups $\mathcal{C} = \{C_1, C_2, \dots, C_K\}$ such that each $C_k \in \mathcal{C}$ describes one semantic topic. The size of the set of groups or the number of semantic topics is assumed to be K . The subset of the training data belonging to a semantic group C_k is denoted by \mathcal{X}_{C_k} . The set of test items is denoted by \mathcal{Y} and each $Y \in \mathcal{Y}$ consists of an image (Y^I), an article (Y^d), a news category label (Y^n), and a keyword (Y^{key}). The textual description of image Y^I is denoted by H_Y and is assumed to be unknown. The set of all news articles is denoted by \mathcal{D} and the set \mathcal{E} contains every named entity e_i mentioned in every news article $d \in \mathcal{D}$.

4.2 News Image Annotation

The news images are an excellent example of a real world set of images with naturally available captions or descriptions. Therefore, they should be widely used for evaluating image annotation system. This has not been the case in past. Most of the image annotation work has been evaluated over image datasets with carefully crafted, artificial annotations which describe the contents of images without any hint to the *context* or the background story of images. Examples of such datasets include MSCOCO, IAPR TC-12, Flickr30K, etc. We discussed the collection of artificial captions for such datasets in Section 4.2.1. Such datasets are a good starting point for evaluation of image annotation work but the annotation system should be matured to deal with real world datasets.

In real world, images are almost always described in their specific *context*. People describe the photos they upload on social media websites in reference to their specific circumstances such as vacation or celebratory events. The images with news articles are described in the context of some news story. Figures 4.1 and 4.2 show sample images from IAPR TC-12 and MSCOCO datasets. Both datasets contain image descriptions manually written by human judges with no knowledge about the *context* of images. Figure 4.3 shows sample images from our TIME dataset of news images. The captions of these images are the actual descriptions used by the news source. Hence, they describe these images in reference to some real world *context*. These figures highlight the difference in nature of artificial and actual image descriptions.

A reasonable image annotation model aims at producing annotations that match the real world ground truth annotations for the input image. If the ground truth annotations include hints to the *context* of the image, it is necessary for the image annotation system to incorporate contextual information. We incorporate the semantic relational information in terms of probability distributions in our model for news image annotation as quantification of the *context* of images.



(a) A big red telephone booth that a man is standing in. (b) A child and woman are cooking in the kitchen. (c) A close-up of a hand touching various pastries.

Figure 4.1: MSCOCO image-caption pairs



(a) A big bunch of flowers with red roses. (b) A meal with meat, onions and rice on a white plate. (c) A man is wearing a grey sweater and a brown hat with red crab on it.

Figure 4.2: IAPR TC-12 image-caption pairs



(a) A flooded street in New Jersey. (b) The National Security Agency headquarters in Fort Meade, Md. (c) On Wall Street, a bull statue.

Figure 4.3: TIME image-caption pairs

4.2.1 Context-sensitive News Image Annotation System

We focus our attention on news images and tailor our semantic contextual relation extraction framework to exploit all contextual information sources available with news images. In addition to the

images, news datasets have quite a few auxiliary information sources, e.g., news articles, structured text such as keywords and news category information, etc. We devise frameworks to extract semantic contextual cues from all of these sources of different data modalities. We employ the probability space as the common ‘representation space’ to combine contextual cues collected from sources of different modalities,. In Sections 3.1 and 3.2 of Chapter 3, we represented the semantic relational information in terms of two probability distributions; 1) the probability distribution of training items conditioned over semantic *themes* or topics, and 2) the probability distribution of semantic topics conditioned over the test image. We extend such probabilistic estimation of the contextual relations to cover all information sources. The goal of our annotation system is to come up with a joint probability distribution $P(\mathbf{r}, \mathbf{w}|\theta_Y)$ of words (\mathbf{w}) and image representation (\mathbf{r}), conditioned over the semantic contextual information (θ_Y) of the test image (Y). This annotation model is part of our manuscript accepted for publication in IEEE Transaction on Image Processing[114].

4.2.1.1 Context Estimation

We have identified four sources of semantic contextual information for news images.

- Semantic scene characteristics of news images
- News articles
- News category labels
- Article keywords

The contextual information collected from each of these source for the test item Y is denoted as the superscript over the context variable θ_Y ($\theta_Y^s, \theta_Y^d, \theta_Y^n, \theta_Y^{key}$ for scene analysis, article, news category and keywords, respectively).

4.2.1.1.1 *Scene Characteristics of Images*

Various studies suggest that while looking at images, humans quickly extract semantic categorical properties of images to identify scenes shown in them, instead of identifying the objects to recognize the scenes[85, 86]. We argue that the scene recognition is not only independent of the object recognition, but it also enables us to make educated guess about the objects in the image. Object detection in an image can be improved by incorporating *context* in the process [96]. Scene analysis of an image can provide the requisite *context* information. We employed the scene characteristics of images to define the semantic categories for our scene-based automatic image annotation model described in Section 3.1 of Chapter 3. We base the *context* estimation for news images from the scene characteristics on a similar process.

Scene recognition is often presented as a classification problem with a finite number of classes or scene-types (‘inside city’, ‘open country’, etc.). It is intuitively evident that the probability distributions over finite set of objects (‘car’, ‘window’, ‘tree’, etc.) would be widely different for images of different scene-types. Such distributions should be used for identifying the image details to be predicted as its textual annotations. A ‘good” method should be biased towards predicting certain annotations, where the bias is based on the scene analysis of the image. We employed two different scene representations for images;

- GIST features describe images in terms of perceptual dimensions (openness, roughness, etc.)[86]. These dimensions can be computed by spectral analysis of images.
- Convolutional neural network have been trained over 2.5 million images of Places dataset to predict approximately 200 scene-types[129]. We processed images through such CNN and extracted image features from the last fully connected layer (fc7) as scene representation vectors for images. We refer to these features as ‘PlacesCNN’.

We devised one framework that can estimate the *context* from any of these scene features. This framework clusters training images (\mathcal{X}) based on the similarity in their scene features (GIST or PlacesCNN). Each cluster \mathcal{X}_{C_s} represents a scene-class or a *context* category C_s . If M_s is the size of \mathcal{X}_{C_s} , then

$$P(X|C_s) = \begin{cases} 1/M_s, & \text{if } X \in \mathcal{X}_{C_s} \\ 0, & \text{otherwise} \end{cases} \quad (4.1)$$

The scene-based *context* information of the test image Y is encoded in its scene features (GIST or PlacesCNN), denoted by θ_Y^s , and

$$P(C_s|\theta_Y^s) = \frac{\exp(-(G_{X_s}^Y - \theta_Y^s)^T \Gamma^{-1} (G_{X_s}^Y - \theta_Y^s))}{\sqrt{2\pi|\Gamma|}}, \quad (4.2)$$

$G_{X_s}^Y$ denotes the scene features of the training image $X^I \in \mathcal{X}_{C_s}$ that is the closest to θ_Y^s . Γ is the covariance matrix, assumed to be of the form $\kappa\mathbf{I}$ for convenience, where \mathbf{I} is the identity matrix, and κ can be selected empirically over a held-out portion of the training dataset.

We employ hierarchical clustering with a cut-off threshold and maximum allowed size of a cluster as the system parameters. If a cluster exceeds the maximum size limit, it is further divided by hierarchical clustering. Clusters with single (or a very small number of) member(s) are dropped. The goal is to come up with a set of *context* clusters such that the size of each cluster falls under a narrow range. This ensures that the training data has a relative even distribution of *context* categories. Note that no supervision is involved in this process and the association of the test image Y^I with *context* categories is not discrete. It is encoded in a continuous domain distribution $P(C_s|\theta_Y^s)$. Hence, the image Y^I is assumed to depict characteristics of multiple types of scenes. The number of clusters dictates the resolution of the scene-based distinction among images.

We thoroughly evaluated the effectiveness of both types of scene features (GIST and PlacesCNN).

We also studied the effects of the number of scene clusters on the performance of our annotation system. Our observations are presented in Section 4.2.2.

4.2.1.1.2 News Articles

News articles discuss various aspects or topics concerning a news story. The image accompanying a certain article is almost always described in reference to the story discussed in the article.



(a) **Caption:** protesters & police in Madrid, Spain.



(b) **Caption:** protests in Moscow

Figure 4.4: Visually similar images - Different news stories.



(a) **Caption:** shuttered Best Buy store in Chicago on April 16, 2012.

Article: BestBuy, low profits, CEO replacement



(b) **Caption:** BestBuy CEO Brian Dunn resigned amid investigation into his 'personal conduct', company said on Tuesday.

Article: BestBuy, low profits, CEO resignation

Figure 4.5: Similar articles - Visually different images

Any annotation system including our systems presented in Chapter 3, can be used to annotate news images. All information from the articles and other available information sources will be ignored. Figure 4.4 shows two images with similar visual contents but different captions. The captions are different from each other because the articles with the two images discuss different news stories from two different countries. It implies that the articles contain invaluable information required to annotate news images with words matching their ground truth captions. Some previously proposed systems discussed in Section 2.1.1 of Chapter 2 ignore visual information of the news images and annotate them only in the light of their articles[66, 67, 80, 20]. Figures 4.5 shows two images which are accompanied by the articles discussing identical sets of topics or the same news story. The two images have vastly different captions based on their visual contents. We argue that the information from visual contents of images and textual contents of the accompanying articles, as well as all other available sources need to be combined to develop better understanding of the data.

We use probabilistic topic modeling to understand the ‘topics’ or the aspects of the news events discussed in news article. Blei et al. proposed latent Dirichlet allocation based generative modeling for document collections[10]. Each document is modeled as a mixture of underlying topics while each topic is represented by a probability distribution over the words of the vocabulary set. Our framework uses similar generative modeling for articles’ collection and considers each topic as a semantic *context* category. The following are the steps involved in such modeling.

- choose $L \sim \text{Poisson}(\eta)$
- choose $\theta_X^d \sim \text{Dir}(\alpha)$
- for each of L words w_l
 - choose topic $C_d \sim \text{Multinomial}(\theta_X^d)$
 - choose a word w_l from $P(w_l|C_d, \xi)$, a multinomial probability conditioned on topic C_d

X^d is the article associated with the training image X^I . θ_X^d is a K -dimensional Dirichlet random variable, where K is the size of the set of underlying topics. These topics form the set of semantic *context* categories, \mathcal{C} , of size K . L is the length of the document which is assumed to be fixed for all the documents. The most interesting estimated distribution is $P(C_d|\theta_X^d)$ which denotes the probability of the topic $C_d \in \mathcal{C}$, conditioned on the article X^d . $P(C_d|\theta_X^d)$ is an estimate of the topics covered in the article X^d and hence, encodes the semantic *context* of the image X^I .

The topics of this modeling scheme are considered the semantic concepts that encode the contextual information necessary to predict meaningful annotations for news images. We employ the article-topic relations as the basis of image-semantic concepts/topics in our framework. In keeping with our general approach of dividing training data into semantic *context* groups, training item X is deemed a member of the semantic *context* group C_d if $P(C_d|\theta_X^d) > 0$, i.e., $\mathcal{X}_{C_d} = \{X | P(C_d|\theta_X^d) > 0\}$. The association of the test item Y with the *context* category or topic C_d is encoded in the distribution $P(C_d|\theta_Y^d)$, estimated by variational inference.

4.2.1.1.3 News Category Labels

Every news media outlet divides the news pieces into a few classes. In this work, we call these classes ‘news categories’. Table 4.1 lists such category labels for various news media outlets. The news categories also provide *context* information for predicting annotations of an image. For example, images in the ‘Business’ category of US-based news papers are more likely to have the words ‘Wall Street’ in their captions than the images from the ‘Entertainment’ category.

Table 4.1 shows that the sets of categories used by different news media outlets are not identical. Still, there are important labels (e.g., ‘politics’, ‘sports’, ‘business’, etc.) which are common among all major news sources. Other labels can be consolidated by careful examination of the news media outlet. For example, the set of news events discussed under ‘Local’, ‘National’, and ‘World’

depends on the origin of news source. U.S. based news sources discuss U.K. related news stories in the ‘World’ or the ‘International’ category while the same stories are discussed under the ‘National’ or the ‘Local’ categories of U.K. based news sources. The New York Times has a dedicated section to discuss the news related to the New York City. The stories discussed under this category may be part of the ‘U.S.’ category for other U.S. based news papers. Hence, the news category labels can be consolidated even when the data is collected from different news sources.

Table 4.1: News categories of a few popular news sources

News Source	Sample of news categories
The Guardian ²	World, Politics, Business, Sports, Tech, US, UK, Lifestyle, Fashion
The Washington Post ³	World, Politics, Business, Sports, Tech, National, Lifestyle, Opinion
The New York Times ⁴	World, Politics, Business, Sports, Tech, U.S., Health, New York, Style

Figure 4.6 shows the distribution of articles across all the news categories for the TIME dataset. Article distribution is very uneven. Some categories contain far larger number of news articles than other categories. This distribution is beyond the control of image annotation system developers.

We argue in favor of exploiting the *context* estimated from the news categories, despite uneven data distribution and the effort needed to consolidate the list of categories used by various news sources. Since the relation between an image and a news category can be an indicator of the words used to describe the image, each news category acts as one semantic topic or *context* category.

We divide the training data into semantic *context* groups based on the news categories assigned to the training items. Since each news category n is treated as one semantic topic/concept, all

²<http://www.theguardian.com/uk>

³<http://www.washingtonpost.com/>

⁴<http://www.nytimes.com/>

items belonging to the category n are the defining members of that semantic context, i.e., $\mathcal{X}_{C_n} = \{X|X^n = n\}$. Since news category labels of test items are also available, the association of the test item Y with the news category based *context* groups is estimated in terms of $P(C_n|\theta_Y^n)$ as

$$P(C_n|\theta_Y^n) = \begin{cases} 1, & \text{if } Y^n = n \\ 0, & \text{otherwise} \end{cases} \quad (4.3)$$

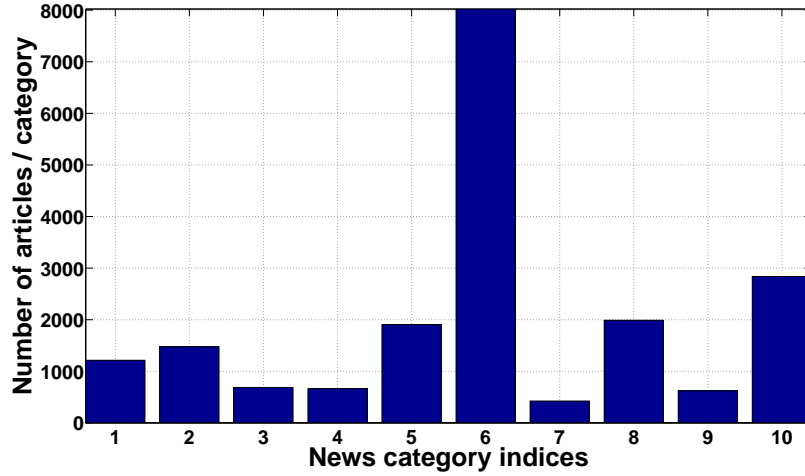


Figure 4.6: Distribution of items among news categories for TIME dataset

4.2.1.1.4 Article Keywords

News papers generally assign one or more keywords to each article. For online editions of news papers, these keywords are usually *hyperlinks* between different articles. The purpose of such keyword assignment is to make it easier for the users reading one article to navigate to other articles that discuss the same news story or similar topics. Hence, these keywords encode information about the topics discussed in the article which provides important semantic contextual information for annotating images associated with articles.

The set of keywords used by any news paper are by nature different from the news category labels used by the same news source. While the set of news categories tends to remain stable for long periods of time, the set of keywords keeps evolving. New keywords are added to the set as new news stories emerge. Old keywords can be re-introduced in a different context. the number of available keywords is usually much larger than the number of news categories for the same news source. Hence, it seems inappropriate to treat keywords in similar fashion as the news categories while extracting semantic contextual relations. Figure 4.7 shows the distribution of articles across the set of keywords for the TIME dataset. It is apparent that the distribution is extremely uneven, implying that some keywords are vastly more popular than the others.

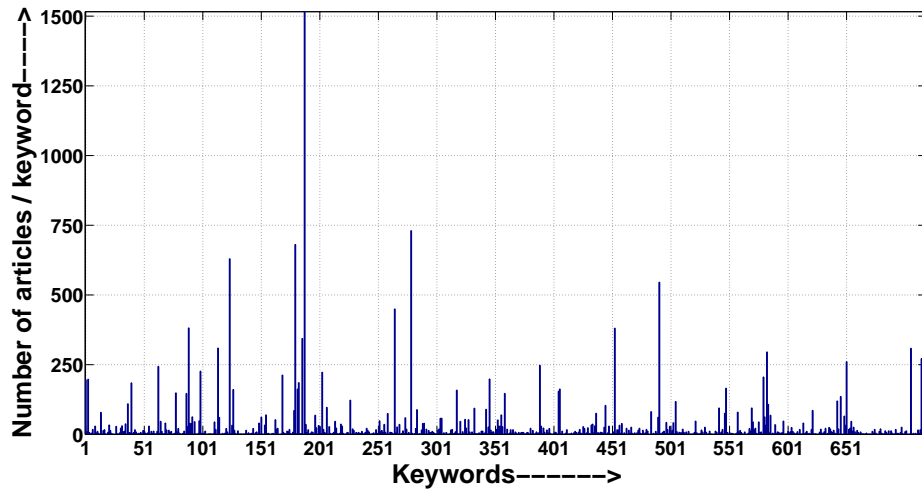


Figure 4.7: Distribution of items among article keywords for TIME dataset

Since keywords are fundamentally used to link articles discussing similar topics together, it is obvious that they hold clues as to what topics are being discussed in the articles. We treat topics discussed in the articles as semantic contextual concepts. We introduced the idea of probabilistic topic modeling to estimate the probability distribution for such concept in Section 4.2.1.1.2. We employ similar technique for probabilistic modeling of topics based on keywords.

Since all articles associated with a certain keyword are assumed to discuss the same set of topics, they can be concatenated to form one *key-document*. There is one *key-document* corresponding to each keyword. Latent Dirichlet allocation based generative modeling is used to model this collection of *key-documents* as a mixture of topics. Topic distribution $P(C_{key}|\theta_X^{key})$ corresponding to each *key-document* is estimated. $P(C_{key}|\theta_X^{key})$ quantifies the association of the keyword *key* with the semantic contextual topic C_{key} . The training items with *key* as their article keyword must have similar association with this semantic topic. All training items with keyword *key* and non-zero $P(C_{key}|\theta_X^{key})$ are the defining members of the semantic *context* group or the topic C_{key} , i.e., $\mathcal{X}_{C_{key}} = \{X | P(C_{key}|\theta_X^{key}) > 0\}$. Keywords for all the test items are also available. Hence, $P(C_{key}|\theta_Y^{key})$ denotes the association of the test item *Y* with semantic contextual topic C_{key} .

4.2.1.1.5 Combination of Heterogeneous Context Sources

We devise a flexible weighted concatenation framework for combining probability distributions which encode semantic contextual cue estimated from various sources. Each source generates a fixed number of semantic concepts and corresponding groups of training items that represent those concepts. Let us assume that three information sources (*g*, *h* and *q*) are being employed, resulting into three sets of semantic *context* categories, i.e., $\mathcal{C}_g, \mathcal{C}_h, \mathcal{C}_q$. $P(C_k|\theta_Y^g)$, $P(C_k|\theta_Y^h)$ and $P(C_k|\theta_Y^q)$ denote the distributions of each of these sets of *context* categories conditioned over the *context* information (θ_Y) of a test item *Y*.

$$P(C_k|\theta_Y) = \begin{cases} \alpha^g P(C_k|\theta_Y^g) & \text{if } C_k \in \mathcal{C}_g \\ \alpha^h P(C_k|\theta_Y^h) & \text{if } C_k \in \mathcal{C}_h \\ \alpha^q P(C_k|\theta_Y^q) & \text{if } C_k \in \mathcal{C}_q \end{cases} \quad (4.4)$$

such that

$$\alpha^g + \alpha^h + \alpha^q = 1 \quad (4.5)$$

The optimal value of $\alpha = [\alpha^g, \alpha^h, \alpha^q]$, tuned with respect to the accuracy of the predicted annotations also indicates the relative quality of the *context* information source. We explore both manual and automatic methods to find the optimal value for vector α .

4.2.1.2 Context-sensitive Generative Model

Here, we present our generative model inspired by relevance models[64], that incorporates both the *context* and the *content* of images for predicting appropriate word annotations for these images. We refer to our annotation model as ‘*context-AIA*’.

Image X^I of the training item $X \in \mathcal{X}$ consists of a set of visual units $\mathbf{r}_X = \{r_{x1}, r_{x2}, \dots, r_{xA}\}$ representing its visual *contents*. X^I is associated with a set of words $\mathbf{w}_X = \{w_{x1}, w_{x2}, \dots, w_{xB}\}$. $P(X|C)$ is the conditional distribution for X over the set of *context* categories. Each word $w_b \in \mathbf{w}_X$ is assumed to be an independent identically distributed (*i.i.d.*) sample from multinomial distribution $P_{\mathcal{W}_{C_k}}(\cdot|X)$. Each visual component is a random sample from multivariate distribution $P_{\mathcal{R}}(\cdot|X)$.

Let Y^I be a new image with its visual *contents* and the *context* encoded in $\mathbf{r}_Y = \{r_{y1}, r_{y2}, \dots, r_{yA}\}$ and θ_Y , respectively. θ_Y consists of four part, i.e., $\theta_Y^s, \theta_Y^d, \theta_Y^n, \theta_Y^{key}$ as described in Section 4.2.1.1. $P(C|\theta_Y)$ is the probability distribution of *context* categories, conditioned over the *context* of the test item Y . Section 4.2.1.1 describes the process of identification of *context* categories C_k in terms of subsets of the training data \mathcal{X}_{C_k} . The estimation of the association of the training and the test items with these categories, i.e., $P(X|C_k)$ and $P(C_k|\theta_Y)$, respectively, is also described in the same section.

$P(\mathbf{r}_Y, \mathbf{w}_Y|\theta_Y)$ needs to be maximized to determine the annotations set \mathbf{w}_Y for the test item Y . The following generative model assumes that the test images and their descriptions are generated

by the same model that generates the training dataset \mathcal{X} . $P(\mathbf{r}, \mathbf{w}|\theta_Y)$ can be estimated through expectation over the training dataset \mathcal{X} .

- pick a context $C_k \in \mathbf{C}$ with probability $P(C_k|\theta_Y)$
 - pick a training image $X \in \mathcal{X}$ with probability $P(X|C_k)$
 - For $b = 1, 2, 3, \dots, B$
 - * pick a word from distribution $P_{\mathcal{W}_{C_k}}(\cdot|X)$
 - For $a = 1, 2, 3, \dots, A$
 - * pick a visual component from distribution $P_{\mathcal{R}}(\cdot|X)$

The above expectation process is summarized as

$$P(\mathbf{w}, \mathbf{r}|\theta_Y) = \sum_{C_k \in \mathbf{C}} P(C_k|\theta_Y) \sum_{X \in \mathcal{X}} P(X|C_k) \prod_{b \in B} P_{\mathcal{W}_{C_k}}(w_b|X) \prod_{a \in A} P_{\mathcal{R}}(r_a|X) \quad (4.6)$$

Each word w_b is an *i.i.d.* sample from $P_{\mathcal{W}_{C_k}}(\cdot|X)$, drawn from a set of words \mathcal{W}_{C_k} . \mathcal{W}_{C_k} is the vocabulary set corresponding to the *context* category C_k . It is a subset of the overall vocabulary set \mathcal{W} ($\mathcal{W}_{C_k} \subseteq \mathcal{W}$). It contains the words used in the descriptions of images of the set \mathcal{X}_{C_k} . $P_{\mathcal{W}_{C_k}}(w_b|X)$ is the w_b^{th} component of this distribution. Its Bayes estimation with Dirichlet prior is

$$P_{\mathcal{W}_{C_k}}(w_b|X) = \frac{\mu \delta_{w_b} + M_{w_b k}}{\mu + M_k} \quad (4.7)$$

δ_{w_b} is 1 if the annotations of X include the word w_b . μ is an empirically selected constant, $M_{w_b k}$ is the number of samples from \mathcal{X}_{C_k} containing the word w_b in their descriptions, and M_k is the size of the set \mathcal{X}_{C_k} .

$P_{\mathcal{R}}(r_a|X)$ is the density estimate for generating the visual component r_a , given a training item X . Assuming that $\mathbf{r}_X = \{r_{x1}, r_{x2}, \dots, r_{xA}\}$ represents visual units of the training image X^I , we employ

the following non-parametric Gaussian kernel based density estimate.

$$P_{\mathcal{R}}(r_a|X) = \frac{\exp(-(r_a - r_{xa})^T \Sigma^{-1} (r_a - r_{xa}))}{\sqrt{2\pi|\Sigma|}} \quad (4.8)$$

The covariance matrix Σ is assumed to be of the form $\beta \mathbf{I}$. \mathbf{I} is the identity matrix. β determines the smoothness around the point r_{xa} and can be determined on a held-out set of the data.

Note that the generative process and the estimation of $P_{\mathcal{W}_{C_k}}(\cdot|X)$ and $P_{\mathcal{R}}(\cdot|X)$ are similar to the ones used in our image annotation model described in Section 3.1.1.2 of Chapter 3.

4.2.2 Evaluation of News Image Annotation System

We used mean precision per word, mean recall per word, and the number of words with positive recall (N^+) as evaluation metrics. We mainly used the grid-based visual features (Section 3.1.1.1 of Chapter 3) for images. Just like the image captions of the dataset like Flickr30K and MSCOCO, ground truth image captions of the TIME dataset are in the form of sentence. Image captions of all of these datasets are tokenized, lemmatized and part-of-speech tagged. Frequently occurring nouns, verbs and adjectives are employed as ground truth annotations. We include much larger vocabulary set in our experiments with the TIME dataset as compared to the vocabulary sets of standard image annotation datasets. Wider visual variety of news images require larger vocabulary set for generation of sufficiently meaningful descriptions.

4.2.2.1 Comparison Models

The availability of the auxiliary information gives rise to various baseline methods such as the use of the titles, the most frequent or the top *tfidf* words of article as annotations. Feng et al.

proposed extended relevance model (extModel)[31] and joint generative modeling of textual and *visual* words (mixLDA)[32] to annotate news image in light of their auxiliary information. We compared the performance of our system against these models. We also included the representatives from two major classes of image annotation models, i.e., multiple Bernoulli relevance model (MBRM) from relevance model based systems, and the TagProp from the nearest-neighbor type systems.

4.2.2.2 Results

Table 4.2 shows the performance of ours and various other frameworks on the BBC dataset which was introduced in Section 4.1. Tables 4.4 and 4.3 contain the performance evaluation results of various baselines and previously proposed models over the TIME dataset, respectively. Tables 4.5 and 4.6 show the evaluation results of our *context*-AIA model over the TIME dataset. The effectiveness of different *context* sources is presented in Table 4.5. The results in Table 4.6 show the effectiveness of different combinations of *context* sources for annotating news images through *context*-AIA model. Our system outperforms various other methods over both datasets TIME and BBC datasets.

4.2.2.3 Observations

The proposed framework induces *context* sensitivity in relevance model and outperforms other relevance model based systems such as MBRM. The nearest-neighbor type algorithms usually perform better than the relevance model based systems, but our *context*-sensitive relevance model based system outperforms signature nearest-neighbor type algorithm, i.e., TagProp. Relatively modest performance of TagProp in our experiments also indicates the inefficiency of the nearest-neighbor approach in dealing with larger vocabulary sets. The size of the vocabulary set used in

our experiments with the TIME dataset is 1200, about four times the size of the sets used for IAPR TC-12 and ESP game datasets used in [39].

Table 4.2: Evaluation of *context*-AIA and comparative annotation models (BBC dataset)

Model	Mean Precision	Mean Recall	Mean F-score
tf-idf[31]	4.37	7.09	5.41
Article Title[31]	9.22	7.03	7.20
CRM (Lavrenko et al., 2003)[31]	9.05	16.01	11.81
txtLDA[32]	7.3	16.9	10.2
CorrLDA[32]	5.33	11.80	7.36
PLSA[32]	10.26	22.60	14.12
extModel[31]	14.72	27.95	19.82
mixLDA[32]	16.3	33.1	21.6
<i>scene-70 & article-100</i>	20	35	25.5

Our framework employs GIST features for the estimation of the semantic *context* from images. We experimented with appending GIST features with the grid-based visual features and using MBRM for image annotation (MBRM-GIST in Table 4.3). The comparison between the scene-based *context* estimation for *context*-AIA and the MBRM-GIST proves that the scene information is better utilized as a source of *context*.

Our method outperforms extModel, txtLDA and mixLDA[31, 32]. Interestingly, mixLDA performs better than the extModel on the TIME dataset. This trend is opposite to that observed on the BBC dataset.

Precision and recall scores show that scene analysis and news categories are two high-quality sources of *context*. Both the GIST and PlacesCNN scene features seem to be equally effective. Therefore, we use only one of these features, i.e., GIST, for further experiments. *Context* from news categories is also the most readily computable. The *context* estimated from the article and

the *keywords* are of modest quality. Combining *context* from multiple information sources helps the performance of the system. When highly effective *context* sources such as *scene* and *category* are combined, precision and recall scores are improved beyond what these sources achieve individually.

Table 4.3: Performance evaluation of previously proposed annotation methods (TIME dataset)

Baseline	Mean Precision	Mean Recall	Mean F-score
Article title	15	11	11
Top tf words	16	15	14
Top <i>tfidf</i> words	13	25	13

Table 4.4: Baseline annotation performance (TIME dataset)

Model	Mean Precision	Mean Recall	N⁺
MBRM	32	15	698
MBRM-GIST	31	16	707
TagProp	15	14	655
extModel	33	15	738
txtLDA	10	9	358
mixLDA	11	19	237

Table 4.5: Comparative performance of *context* sources of *context*-AIA (TIME dataset)

<i>context</i>-AIA configuration	Mean Precision	Mean Recall	N⁺
<i>scene</i> -Places-300	44	21	753
<i>scene</i> -100	44	21	758
<i>scene</i> -50	43	20	742
<i>scene</i> -20	38	20	699
<i>article</i> -1000	21	9	479
<i>article</i> -500	22	10	508
<i>article</i> -100	23	10	517
<i>category</i> -10	40	20	717
<i>key</i> -1000	24	11	541
<i>key</i> -500	22	10	526
<i>key</i> -100	23	10	529

Table 4.6: Performance of combinations of *context* sources of *context*-AIA (TIME dataset)

<i>context</i>-AIA configuration	Mean Precision	Mean Recall	N⁺
<i>scene</i> -100 & <i>article</i> -1000	44	20	748
<i>scene</i> -100 & <i>article</i> -500	44	20	749
<i>scene</i> -100 & <i>article</i> -100	44	20	750
<i>scene</i> -50 & <i>article</i> -1000	43	20	737
<i>scene</i> -50 & <i>article</i> -500	43	20	737
<i>scene</i> -50 & <i>article</i> -100	42	20	739
<i>scene</i> -20 & <i>article</i> -1000	40	19	696
<i>scene</i> -20 & <i>article</i> -500	40	19	694
<i>scene</i> -20 & <i>article</i> -100	40	19	693
<i>scene</i> -100 & <i>category</i> -10	44	21	748
<i>scene</i> -50 & <i>category</i> -10	45	21	742
<i>scene</i> -20 & <i>category</i> -10	38	20	702
<i>category</i> -10 & <i>article</i> -1000	39	19	682
<i>category</i> -10 & <i>article</i> -500	39	19	681
<i>category</i> -10 & <i>article</i> -100	38	19	682
<i>scene</i> -100 & <i>key</i> -1000	44	20	751
<i>scene</i> -100 & <i>key</i> -500	44	20	750
<i>scene</i> -100 & <i>key</i> -100	44	20	749
<i>scene</i> -50 & <i>key</i> -1000	44	20	736
<i>scene</i> -50 & <i>key</i> -500	43	20	739
<i>scene</i> -50 & <i>key</i> -100	43	20	739
<i>scene</i> -20 & <i>key</i> -1000	39	20	696
<i>scene</i> -20 & <i>key</i> -500	40	20	695
<i>scene</i> -20 & <i>key</i> -100	40	20	694
<i>category</i> -10 & <i>key</i> -1000	39	19	682
<i>category</i> -10 & <i>key</i> -500	39	19	681
<i>category</i> -10 & <i>key</i> -100	38	19	682
<i>scene</i> -50 & <i>category</i> -100 & <i>article</i> -100 & <i>key</i> -1000	46	21	737

News articles contain rich semantic information but perform only modestly well when introduced as a *context* source. This trend can be explained in terms of the noise. News article may discuss a number of different ‘topics’ while the accompanying image may be relevant to only a few of them. Therefore, when all ‘topics’ of the news article are used as *context*-cues, the system faces difficulty

in focusing on the ‘topics’ most relevant to the image contents. The evolving nature of *keywords* explain the relative inefficiency of news *keywords* as a *context* source for image annotation.

We also experimented with the number of *context* categories, generated from each source (except for the news category as the number of news categories is fixed for the dataset). For the scene analysis, the number of *context* categories is controlled by the cut-off threshold and the maximum allowed size of cluster used in hierarchical clustering. For article and *keywords*, the number of topics in topic modeling process decides the number of generated *context* categories. The performance remains stable for a wide range of the number of *context* categories. Increasing the number of categories beyond a certain limit, does not improve the performance.

The vocabulary varies from one *context* category to another (\mathcal{W}_{C_k} for *context* category C_k) instead of being fixed to a specific number for all the data, in *context*-AIA model. We ensured that approximately the same number of unique words appear in the final output, i.e., the annotations predicted for the test images by adjusting the parameters of our method. The number was fixed to 1200 for the TIME dataset. Thus the results of different versions of our method are comparable to each other and to the baseline results.

4.2.2.4 Parameter optimization

A significant quality of our *context* estimation strategy is that the information from heterogeneous *context* sources can be combined to take advantage of all the available information. This process requires weighted concatenation of $P(C_k|\theta_Y)$ estimated from different sources of *context* information. As described in Section 4.2.1.1, the weight vector α optimized with respect to the accuracy of predicted annotations. In addition to optimizing the performance of the annotation system, the entries of the optimal weight vector α indicate the effectiveness of different *context* sources for predicting annotations for news images. We explored two methods to optimize α .

4.2.2.4.1 Manual Tuning

We used a validation set, consisting of randomly picked 10% items from the training data, as held-out dataset for tuning the weight vector α with respect to the accuracy of the predicted annotations. This approach was used to generate the results reported in Section 4.2.2.2.

4.2.2.4.2 Least Squares Error Minimization

The estimation of vector α can be modeled as a *least squares error* (LSE) minimization problem. Assuming that three *context* sources are being used,

$$\alpha = [\alpha^1, \alpha^2, \alpha^3] \quad (4.9)$$

Let \mathcal{X}' denote the validation dataset where $M' = |\mathcal{X}'|$. Vector \mathbf{u}^g contains the values of $P(w, \mathbf{r}'_{\mathbf{X}} | \theta_{X'}^g)$ such that $X' \in \mathcal{X}'$ and $w \in \mathcal{W}$, estimated by employing only the g_{th} *context* source. If $N = |\mathcal{W}|$, then the length of \mathbf{u}^g is $N \times M'$. Indices $[(m' - 1) \times N + 1, m' \times N]$ of the vector \mathbf{u}^g contain the joint probability estimates for m'_{th} image where $m' \in [1, M']$. If $p_{nm'}^g$ denotes the joint probability estimate of the n_{th} word with the m'_{th} image with g_{th} *context* source, then

$$\mathbf{u}^g = [p_{11}^g, p_{21}^g, \dots, p_{N1}^g, p_{12}^g, p_{22}^g, \dots, p_{N2}^g, p_{1M'}^g, p_{2M'}^g, \dots, p_{NM'}^g] \quad (4.10)$$

One such vector is generated from our *context*-AIA annotation model for every available *context* source. Weighted combination of information from all sources translates to the weighted sum of the corresponding $p_{nm'}^g$ from all of these vectors. If $g \in \{1, 2, 3\}$, then

$$\mathbf{u} = \alpha^1 \times \mathbf{u}^1 + \alpha^2 \times \mathbf{u}^2 + \alpha^3 \times \mathbf{u}^3 \quad (4.11)$$

$$\mathbf{u} = [\alpha^1, \alpha^2, \alpha^3] \times \begin{pmatrix} \mathbf{u}^1 \\ \mathbf{u}^2 \\ \mathbf{u}^3 \end{pmatrix} \quad (4.12)$$

The ground truth \mathbf{g} is a binary vector of size $N \times M'$. Entries from the indices $[(m' - 1) \times N) + 1, m' \times N]$ of \mathbf{g} represent the presence or the absence of each word of the vocabulary in the description of the m'_{th} image from the set \mathcal{X}' where $m' \in [1, M']$. Ground truth information for the set \mathcal{X}' is available as it is a subset of the labeled training dataset. The goal is to minimize the distance between the vectors \mathbf{u} and \mathbf{g} , i.e., to minimize the error in the annotation prediction. The comparison between the vectors \mathbf{u} and \mathbf{g} poses a problem. Vector \mathbf{u} contains joint probability estimate for words and images, and is in continuous domain with range $[0, 1]$. \mathbf{u} is binary vector with 0/1 values.

The output of the annotation model is a continuous-domain joint probability over all images and words. Each image is tagged with a certain number (say B) of words. Each image is assigned a vector of length N such that only the entries corresponding to the top B words, according to the the joint probability estimate, are set to 1. This is the binary output.

The system needs to know the weight vector α before converting the continuous domain system output (\mathbf{u}) to the binary domain for it to be compared against the ground truth represented by the binary vector \mathbf{g} . Calculating the optimal α vector is the goal of this optimization process. Let $\hat{\mathbf{u}}$ be the binary vector obtained after discretizing vector \mathbf{u} and Ψ denote the discretization operator

$$\hat{\mathbf{u}} = \Psi(\alpha \times \mathbf{u}) \quad (4.13)$$

Ψ is clearly a non-linear operator with α as its operand, prohibiting optimization of α . To tackle this problem, we introduce *soft-max* based approximation of Ψ .

Soft-max is a normalized exponential function that is a smoothed approximation of the *max* function.

$$\sigma(z)_f = \frac{\exp(z_f)}{\sum_e \exp(z_e)} \quad (4.14)$$

It approximates the *max* function since for $z_f \gg z_e$ for all $e \neq f$, $\sigma(z)_f \simeq 1$ and $\sigma(z)_e \simeq 0$ for all $e \neq f$ [7]. To control the decay of this exponential function, \mathbf{z} is set as a weighted version of the original vector \mathbf{t} , i.e., $\mathbf{z} = \nu \mathbf{t}$. Higher value of ν implies sharper or more steep decay of the *soft-max* function.

Soft-mas with a reasonable value of ν can approximate the operator Ψ . Value of ν is selected to ensure that the decay of the function allows for the top few entries of the vector to have substantial non-zero values, while making other entries close to zero. Let $\bar{\Psi}$ be the *soft-max* inspired approximation of Ψ operator, then

$$\bar{\mathbf{u}} = \bar{\Psi}(\boldsymbol{\alpha} \times \mathbf{u}) \quad (4.15)$$

$$\bar{\mathbf{u}} = \bar{\Psi}(\alpha^1 \mathbf{u}^1 + \alpha^2 \mathbf{u}^2 + \alpha^3 \mathbf{u}^3) \quad (4.16)$$

$\bar{\Psi}$ is still a non-linear function as it includes an exponential. For the ease of computation, we assume that our method operates in a limited region of the range of $\bar{\Psi}$, where the response of the function is approximately linear. This implies that we can modify Equation (4.16) as

$$\bar{\mathbf{u}} = \alpha^1 \bar{\Psi}(\mathbf{u}^1) + \alpha^2 \bar{\Psi}(\mathbf{u}^2) + \alpha^3 \bar{\Psi}(\mathbf{u}^3) \quad (4.17)$$

$$\bar{\mathbf{u}} = [\alpha^1, \alpha^2, \alpha^3] \times \begin{pmatrix} \bar{\Psi}(\mathbf{u}^1) \\ \bar{\Psi}(\mathbf{u}^2) \\ \bar{\Psi}(\mathbf{u}^3) \end{pmatrix} \quad (4.18)$$

If

$$\Omega = \begin{pmatrix} \bar{\Psi}(\mathbf{u}^1) \\ \bar{\Psi}(\mathbf{u}^2) \\ \bar{\Psi}(\mathbf{u}^3) \end{pmatrix} \quad (4.19)$$

then

$$\bar{\mathbf{u}} = \boldsymbol{\alpha} \times \Omega \quad (4.20)$$

Vectors $\bar{\Psi}$ and \mathbf{g} are compared. In the optimal situation, the entries in the $\bar{\Psi}$ corresponding to 1's in the vector \mathbf{g} have much higher values than the entries corresponding to the 0 entries of \mathbf{g} . We make use of the LSE formulation to minimize the squared distance between $\bar{\Psi}$ and \mathbf{g} . Ideally, $\bar{\Psi} = \mathbf{g}$, and from Equation 4.20

$$\boldsymbol{\alpha} \times \Omega = \mathbf{g} \quad (4.21)$$

$$\boldsymbol{\alpha} \Omega \Omega^T = \mathbf{g} \Omega^T \quad (4.22)$$

$$\boldsymbol{\alpha} (\Omega \Omega^T) (\Omega \Omega^T)^{-1} = \mathbf{g} \Omega^T (\Omega \Omega^T)^{-1} \quad (4.23)$$

Since $(\Omega \Omega^T) (\Omega \Omega^T)^{-1} = \mathbf{I}$ where \mathbf{I} is the identity matrix, $\boldsymbol{\alpha}$ is calculated as

$$\boldsymbol{\alpha} = \mathbf{g} \Omega^T (\Omega \Omega^T)^{-1} \quad (4.24)$$

We observed that the ratios between the optimal weights for different *context* sources of the TIME dataset, through manual tuning and through the LSE optimization are proportional to each other. Table 4.7 presents the optimal weights for the *context* sources, optimized though LSE, in terms of ratios to the optimal weight of the *context* from scene analysis. For example, the first row indicates that the optimal weight for the article based *context* is 15% of that of the scene based *context*.

The *context* estimated from the scene analysis is of the highest quality as all other *context* sources are assigned fractions of its weight under the optimal conditions. News category is also a high quality *context* source as its optimal weight is 73% of that of the scene based *context*. It is consistent with the observations made in the Section 4.2.2.3. As discussed in Section 4.2.1.1, a comprehensive list of news categories needs to be generated if the data is collected from multiple news source. The quality of the *context* estimated from the news categories warrants this additional effort. The optimal weight for the *context* from news articles is modest (15% of the scene-based *context*) while that for keywords is extremely low. As discussed in Section 4.2.2.2, the noise introduced through the news articles and evolving nature of the keywords can explain these weights.

Table 4.7: Comparative quality of *context* source

<i>Context source</i>	Optimal weight ratio	Base <i>context</i> source
Topic modeling of article	0.15	Scene Analysis
Metadata: News Category	0.73	Scene Analysis
Metadata: <i>keywords</i>	≈ 0	Scene Analysis

4.3 News Image Caption Generation

Complex language modeling schemes have been previously proposed to generate sentence-like captions for images[61, 83, 126, 118, 33]. In recent past, sequential neural networks have gained tremendous popularity for generation of word sequences or sentences to describe images[79, 25, 121, 125, 56, 54]. All of these schemes have been proposed for images with no auxiliary information source available. Sentence generation through any form of modeling is quite time consuming. The resulting sequence may still have grammatical errors. Feng et al. evaluated their sentence modeling scheme in terms of grammatical correctness. Their reported results are not satisfactory[33].

News images have vast amount of text already associated with them, in the form of news articles. When an image is associated with the article, it is reasonable to assume that the image is described in at least some portion of the article. We model news image caption generation as the *extraction* of the portion of text describing the image from the article associated with the image. Our framework is somewhat similar to the headline generation or *extractive* document summarization systems. Text summarization techniques can be broadly classified into two classes: *a) Extractive* and *b) Abstractive* [52, 78]. In a nutshell, *extractive* techniques rely on selecting the best sentences from the available text, while *abstractive* techniques try to put together a sentence with the help of language models. We develop a framework that can *extract* the relevant sentence from the article as image caption, instead of building a sentence from scratch. Like our annotation model, our caption generation framework is also *context-sensitive*. We call our framework ‘*context-EXT*’ and it is part of our manuscript accepted for publication in IEEE Transaction on Image Processing[114].

4.3.1 Context-sensitive News Image Caption Generation System

Our caption generation framework estimates a probability distribution over words for each image that the appropriate image caption should have, conditioned over all the semantic contextual cues available for the image. This probability distribution is denoted by $P(w \in H_Y | Y, \theta_Y)$ where H_Y denotes the appropriate caption for the image Y^I of the test item Y . Such estimation provides a template for the sentence that can describe the image properly. Hence, this distribution $P(w \in H_Y | Y, \theta_Y)$ can be used as the matching criterion while searching through an article for the best sentence that can describe the image associated with that article. There are two important information sources while estimating this probability distribution.

- Image
- Article

4.3.1.1 Context-sensitive Word Distribution from Image

The influence of image Y^I over the probability distribution $P(w \in H_Y|Y, \theta_Y)$ is encoded in the distribution $P(w \in H_Y|Y^I, \theta_Y)$ which can be estimated easily through our image annotation framework *context-AIA*. *Context-AIA* estimates $P(\mathbf{r}, \mathbf{w}|\theta_Y)$, i.e., the joint probability of words and visual units conditioned over the *context* of the test image. $P(\mathbf{r}_Y, \mathbf{w}|\theta_Y)$ encodes the influence of image Y^I on the appropriate word distribution $P(w \in H_Y|Y, \theta_Y)$ for the image caption.

$$P(w \in H_Y|Y^I, \theta_Y) \propto P(\mathbf{r}_Y, w|\theta_Y) \quad (4.25)$$

Since, the distribution $P(\mathbf{r}_Y, \mathbf{w}|\theta_Y)$ is dependent over the *context* of the image Y^I , it ensures, in turn, that the distribution $P(w \in H_Y|Y^I, \theta_Y)$ is sensitive to the image's *context*.

4.3.1.2 Context-sensitive Word Distribution from Article

The influence of article Y^d over the word distribution $P(w \in H_Y|Y, \theta_Y)$ is encoded in the distribution $P(w \in H_Y|Y^d, \theta_Y)$. The distribution $P(w \in H_Y|Y^d, \theta_Y)$ quantifies the probability of a word w being part of the caption conditioned over the article and the *context* of an image.

The distribution $P(w \in H_Y|Y^d, \theta_Y)$ can be estimated through a weighted expectation process over the training data as the training data consists of images that have both their articles and captions available.

$$P(w \in H_Y|Y^d, \theta_Y) = P(w \in Y^d) \sum_{C_k \in \mathcal{C}} P(C_k|\theta_Y) P(w \in H_X|w \in X^d, X \in \mathcal{X}_{C_k}), \quad (4.26)$$

$P(w \in Y^d)$ is a simple indicator of whether or not the word w is present in the article Y^d associated with the test image Y^I . $P(w \in H_X|w \in X^d, X \in \mathcal{X}_{C_k})$ is the probability of word w being in the

caption if w is present in the corresponding article for training items of a certain *context*-category C_k . $P(C_k|\theta_Y)$ is the probability of selection of a certain *context*-category C_k conditioned over the *context* information of the test image. Hence, the influence of the article over the word distribution of the caption is also estimated in a *context*-sensitive fashion.

4.3.1.3 Extraction of Caption

Our caption generation framework *context*-EXT combines the influence of the image and the article over the word distribution of the appropriate caption, i.e., $P(w \in H_Y|Y, \theta_Y)$, in a flexible fashion through weighted aggregation.

$$P(w \in H_Y|Y, \theta_Y) = \phi P(w \in H_Y|Y^I, \theta_Y) + (1 - \phi) P(w \in H_Y|Y^d, \theta_Y) \quad (4.27)$$

Constant ϕ denotes the relative emphasis on two contributing distributions representing information from the image and the article. This weighted aggregation scheme gives our model the flexibility to put more emphasis on any of the two information sources if the two sources seem to have uneven influence over the selection of appropriate image caption. $\phi = 1$ presents the case when the caption H_Y is predicted only on the basis of the image Y^I . $\phi = 0$ present the case when the H_Y is predicted only on the basis of the article Y^d . Our experiments show that the best performance is achieved when $0 < \phi < 1$, i.e., the information from both the image Y^I and the article Y^d is combined.

As explained earlier, *Context*-EXT *extracts* the best sentence from the article Y^d associated with the image Y^I to be used as caption H_Y . Let us denote a sentence from article Y^d as vector s_{Y^d} . s_{Y^d} is an N -dimensional vector where N is the size of the vocabulary set \mathcal{W} . n^{th} entry of s_{Y^d} indicates the frequency of the word w_n is the corresponding sentence.

The system iterates through all sentence of the article Y^d and selects the sentence that matches the distribution $P(w \in H_Y|Y, \theta_Y)$ most closely. Our framework employs *cosine*-similarity as the matching criterion. The selected sentence is used as the caption H_Y for image Y^I .

$$H_Y = \underset{s_{Y^d}}{\operatorname{argmax}} \{ \operatorname{cosine}(s_{Y^d}, P(w \in H_Y|Y, \theta_Y)) \} \quad (4.28)$$

4.3.2 Evaluation of News Image Caption Generation System

The goal of any caption generation system is to produce sentences given an image, that describe the image in an appropriate fashion. The sentences should also be grammatically correct so that they make sense to the readers.

Human judges can be asked to go through the image-caption pairs and judge the quality of the generated captions in terms of their relevance to the images as well as grammatical correctness. Such manual evaluation has been previously employed in [33]. There are certain disadvantages to such evaluation. Human judgment is subjective. Two judges may not agree on what should be deemed relevant to the image. Such evaluation scheme is also tedious to execute and expensive in terms of the effort and the time. In case of news images, judges will have to take into account the *context* of the image as well. This may well increase the level of subjectivity in evaluation.

An automatic evaluation scheme is extremely beneficial as it solves the problem of subjectivity as well as being less time-consuming. Evaluation criterion from machine translation have been recently used for evaluating system-generated captions[33, 54, 45, 118, 63]. For machine translation systems, the system-generated translation and the reference translation are compared against each other. In case of caption generation systems, the system-generated caption is compared against the ground truth caption assuming that the ground truth caption is available. Various evaluation measures have been proposed in the past to evaluate machine translation systems.

BLEU is the product of an n-gram precision score and a brevity penalty (BP). BP ensures that the lengths of the system-generated sentences are comparable to those of reference sentences[92].

$$BLEU = BP \times \exp\left(\sum_{n=1}^N \omega_n \log(P_n)\right) \quad (4.29)$$

P_n denotes the n-gram precision. Weights ω_n sum to 1. If N_c and N_r denote the lengths of the test and the reference sentence, respectively, then

$$BP = \begin{cases} 1, & \text{if } N_c > N_r \\ \exp(1 - \frac{N_r}{N_c}), & \text{if } N_c \leq N_r \end{cases} \quad (4.30)$$

METEOR includes unigram matching as well as more advanced matching forms, e.g., paraphrase, stemmed word or synonym matches. As a result, METEOR achieves very high correlation with human evaluation [4].

$$METEOR = F_{mean} \times (1 - penalty) \quad (4.31)$$

where F_{mean} is harmonic mean of unigram based precision and recall and

$$penalty = 0.5 \times \left(\frac{\#chunks}{\#unigram_matches}\right)^3 \quad (4.32)$$

TER is the count of edits required to transform the system-generated sentence to match the reference sentence[87].

$$TER = \frac{INS + DEL + SUB + SHFT}{N_r} \quad (4.33)$$

where INS, DEL, SUB, SHFT indicate the numbers of insertions, deletions, substitutions and shifts required to transform the sentence. N_r denotes the length of the reference sentence. Lower TER score indicates higher translation quality. There is no provision to reward comparable lengths of

the hypothesis and the reference sentences. Thus, shorter sentences tend to score lower on TER.

Our framework *extracts* a sentence from a well-written news article. It is bound to produce grammatically correct captions. The main quality to evaluate is the relevance of the caption to the image and its *context*. Availability of ground truth captions with the news images enables us to use machine translation quality measure like TER, METEOR and BLEU for evaluation.

4.3.2.1 Comparison Models

Since an article is available with each news image in the TIME datasets, we devised one simple baseline model by using the title of the articles associated with images as their captions. This baseline does not require any significant processing.

To estimate the benefits of *context* in caption generation procedure, we devised a second baseline called *baseline-EXT*. This baseline ignores the *context* of the image. It employs *cosine* matching between the sentences of the article and $P(w \in H_Y|Y)$ to *extract* the caption from the article. $P(w \in H_Y|Y)$ is the word distribution without the *context* information θ_Y .

$$H_Y = \underset{s_Y^d}{\operatorname{argmax}} \{ \cos(\cosine(s_Y^d, P(w \in H_Y|Y))) \} \quad (4.34)$$

Distribution $P(w \in H_Y|Y)$ is conditioned over only the test item Y and not its *context* θ_Y . We employ joint probability of words w and visual features r from annotation models which are insensitive to the *context* θ_Y , to estimate the distribution $P(w \in H_Y|Y)$.

$$P(w \in H_Y|Y) \propto P(w, \mathbf{r}_Y) \quad (4.35)$$

We selected one representative from every major classes of annotation systems such that the cho-

sen representative produces highly accurate word annotations for the TIME dataset (Table 4.3), i.e., MBRM (relevance model based annotation systems), TagProp (the nearest-neighbor type annotation systems), extModel (auxiliary information dependent annotation models). Model *baseline-EXT* is insensitive to the *context* when it employs joint probability of the words and the visual features estimated through the annotation models which are insensitive to the *context* of images.

Phrase-based *abstractive* caption generation approach (*phrase-ABS*) [33] is described as

$$\begin{aligned}
P(p_1, p_2, \dots, p_J) &\approx \prod_{j=1}^J P(p_j \in H_Y | p_j \in Y^d) \cdot \prod_{j=2}^J P(p_j | p_{j-1}) \cdot P(\text{length}(H_Y)) \\
&= \sum_{j=1}^J \text{length}(p_j) \cdot \prod_{i=3}^{\sum_{j=1}^J \text{length}(p_j)} P_{\text{adap}}(w_i | w_{i-1}, w_{i-2}) \quad (4.36)
\end{aligned}$$

Phrase, p_j , is a *head* (limited to the types of nouns, verbs and prepositions) with its *modifiers*.

$$P(p_j \in H_Y | p_j \in Y^d) = \prod_{w_j \in p_j} P(w_j \in H_Y | w_j \in Y^d) \quad (4.37)$$

$$P(p_j | p_i) = \frac{1}{2} \sum_{w_i \in p_i} \sum_{w_j \in p_j} \left\{ \frac{f(w_i, w_j)}{f(w_i, -)} + \frac{f(w_i, w_j)}{f(-, w_j)} \right\} \quad (4.38)$$

$P(p_j | p_i)$ is the phrase attachment probability. $f(w_i, w_j)$ is the number of times two phrases containing words w_i and w_j are adjacent. $P_{\text{adap}}(w_i | w_{i-1}, w_{i-2})$ is the trigram model adapted to the probabilities of the annotation system[33]. *phrase-ABS* is computationally more expensive than *context-EXT*. We employed $P(\mathbf{w}, \mathbf{r}_Y)$ from MBRM, TagProp, extModel and *context-AIA* to adapt the distribution $P_{\text{adap}}(w_i | w_{i-1}, w_{i-2})$. *phrase-ABS* is generally insensitive to the *context*, except when $P(\mathbf{w}, \mathbf{r}_Y | \theta_Y)$ from our *context-AIA* model is incorporated in its processing. In that case, the evidence for caption from the image itself is estimated in a *context*-sensitive manner through our annotation model, while the information from the article is processed without the *context*.

4.3.2.2 Results

Tables 4.8 and 4.9 show the comparative performance of different caption generation methods over the TIME and the BBC datasets, respectively (the best scores are indicated by bold font).

4.3.2.3 Observations

Table 4.9 shows performance trends in terms of METEOR and TER on TIME dataset. The proposed *context*-EXT performs the best among all the tested methods in terms of METEOR. The captions generated by titles and phrase-based *abstractive* technique (*phrase*-ABS) are of significantly smaller length as compared to all other methods. By design, TER scores for these methods are better than for all other techniques. Even in this case, the best score is generated by using $P(\mathbf{w}, \mathbf{r}_Y | \theta_Y)$ from our *context*-sensitive annotation model *context*-AIA in phrase-based *abstractive* technique *phrase*-ABS (denoted by *phrase*-ABS(*context*-AIA)). This result further proves the significance of the *context* of the image in the process of generating its caption.

We also compared *context*-EXT for the BBC dataset in terms of TER scores. As described earlier, TER scores are dependent on the average length of the system-generated captions. We use this measure because Feng et al. presented their experimental results as TER scores for the BBC dataset[33]. Table 4.8 is divided in two blocks such that each block contains methods generating captions of similar average length. First block contains baseline models and *extractive* techniques. Our *context*-EXT model achieves the best score among the models included in this block. In the second block, word and phrase based *abstractive* techniques are compared against each other. The best TER score is achieved by the phrase based *abstractive* technique when it employs the probability estimated from our *context*-sensitive annotation model (denoted by *phrase*-ABS(*context*-AIA)). This result is consistent with our observations made over the results for the TIME dataset.

Table 4.8: Performance of caption generation systems over BBC dataset (Average length of ground truth caption is 10); The best score in each block is indicated by bold font.

Model	Approx. avg. length of captions	TER
Lead sentence[33]	21	2.12
Word overlap[33]	24	2.46
<i>baseline</i> -EXT(cosine[33])	22	2.26
KL-divergence[33]	18	1.77
JS-divergence[33]	19	1.77
<i>context</i>-EXT	20	1.75
<i>word</i> -ABS	10	1.11
<i>phrase</i> -ABS(extModel)	10	1.06
<i>phrase</i>-ABS(<i>context</i>-AIA)	9	1.04

Table 4.9: Performance of caption generation systems over TIME dataset. Average length of ground truth caption is 20; The best score in each column is indicated by bold font. [†]:Significantly different from *context*-EXT in terms of METEOR. *:Significantly different from *phrase*-ABS(*context*-AIA) in terms of METEOR.

Model	Avg. Length of caption	METEOR	TER
Titles ^{†*}	10	0.038	1.04
<i>phrase</i> -ABS(MBRM) [†]	7	0.011	1.032
<i>phrase</i> -ABS(TagProp) ^{†*}	6	0.01	1.033
<i>phrase</i> -ABS(extModel) [†]	7	0.007	1.05
<i>phrase</i>-ABS(<i>context</i>-AIA)[†]	7	0.013	1.031
<i>baseline</i> -EXT(MBRM) ^{†*}	21	0.047	1.33
<i>baseline</i> -EXT(TagProp) ^{†*}	21	0.043	1.35
<i>baseline</i> -EXT(extModel) ^{†*}	16	0.034	1.23
<i>context</i>-EXT*	21	0.053	1.32

We ran Wilcoxon test to verify that the performance difference between *context*-sensitive caption generation techniques and methods that ignore *context* is statistically significant. We observed that the difference between METEOR scores of *context*-sensitive strategies (*context*-EXT and *phrase*-ABS(*context*-AIA)), and most other methods is statistically significant at default significance level of 0.05 (Table 4.9).

Table 4.10: Characteristics of *context*-EXT framework; The underlined words in system-generated captions overlap with corresponding ground truth captions.









Image				
Article Summary	A report on Facebook’s advertising model ahead of its IPO.	A report on Micheal Dell’s buyout plan for Dell Inc.	A lawsuit filed by Detroit city attorney was struck down by a judge, Detroit mayor Dave Bing has vocally objected to the lawsuit.	J.K. Rowling’s new book, titled ‘Fantastic Beasts and Where to Find Them’
Ground truth Caption	Facebook CEO Mark Zuckerberg	Michael Dell, chairman and CEO of Dell Inc.	Detroit Mayor Dave Bing	Author JK Rowling
context-EXT	<u>Facebook</u> was originally not created to be a company, <u>Zuckerberg</u> wrote in Facebook’s prospectus.	<u>Michael Dell</u> can breathe easy after Thursday, its unlikely that the barbarians will be at the gate of his namesake PC maker.	He said the suit would have to have been filed by <u>Mayor Dave Bing</u> or the city council.	<u>Rowling</u> has also mentioned that the story starts in New York City; that and the date evoke the delicious possibility of some American-style Jazz-age wizardry.
Image				
Article Summary	Epic Records announced the long awaited Fiona Apple album.	Backlash and lawsuits concerning Dukan diet by Dr. Pierre Dukan.	A report on website ‘40 Days of Dating’	Lawsuit filed by Karen Feld against her brother
Ground truth Caption	Fiona Apple	Dr. Pierre Dukan	Jessica Walsh and Timothy Goodman	Karen Feld and her dog Campari are seen at her home in Washington.
context-EXT	Apple, who hasn’t released any music since her 2005 album Extraordinary Machine, has been rumored to have an album’s worth of new material for quite sometime.	If found guilty, the BBC reports that <u>Dr. Dukan</u> could be removed from the French medical registry.	Still, its an accurate description of what transpires on 40 Days between graphic designers <u>Jessica Walsh and Timothy Goodman</u> .	A Washington jury rejected both <u>Karen Feld’s</u> claim of assault and her brother Kenneth’s counter-claim of trespassing.

Table 4.11: Comparative analysis of captions generated by various systems

Image				
Article Summary	Musical endeavors of Jewish reggae singer Matisyahu	Jason Ready, a white supremacist.	Daniel Radcliffe said that there would no more HP movies.	Lun Lun , a 15 year old panda, gave birth to twins at Atlanta Zoo.
Ground truth Caption	Matisyahu performs during the TEN featuring An Acoustic Evening with Matisyahu at ...	Jason 'JT' Ready	British author J.K. Rowling poses with a copy of her new book 'Harry Potter ...	Lun Lun giving birth two her newborn twins at Atlanta Zoo.
baseline-EXT (MBRM)	The new track Cross-roads, which you can hear exclusively on TIME.	Lilly was taken to a nearby hospital where she was pronounced dead.	Head to Techland for more.	The <u>zoo</u> plans to wait a few months before the cubs are officially named, so hang tight.
phrase-ABS	follow say going came name had	say want the many things a woman is regal	follow say tell have the authors	follow look know didn't weighing
context-EXT	Matisyahu is currently on world tour, with upcoming dates in Los Angeles, ...	Ready was also a prospective candidate for Pinal County sheriff. Harry in the, Harry Potter movies, has said J.K. Rowling told him he would never haveevent in panda history, Lun Lun, a 15 year-old <u>giant panda</u> , gave birth to twins at <u>Zoo Atlanta</u> .
Image				
Article Summary	High Roller USA developed an adult-sized low-riding plastic trike.	Wade Michael Page fired on worshipers at a Sikh temple in Milwaukee.	A team of explorers discovered gold coins off the Florida coast.	A report on Occupy Wall Street gathering in New York.
Ground truth Caption	Your childhood ... in adult size.	A police K-9 unit, left, and a robot, center, take their places outside the Sikh temple in Oak Creek, Wis., where a shooting ...	A shipwreck salvage company recently found these gold coins, known as escudos, just 100 feet off the Florida coast.	Demonstrators with 'Occupy Wall Street' occupy Zuccotti Park on September 29, 2011 in New York.
baseline-EXT (MBRM)	High Roller has taken over my life Arbruster told USA Today.	FBI officials said at a news conference that weapons had been found at the scene.	Days like these are not nearly as common and make all those hard miserable days worth it.	It would be very easy to bring people together.
phrase-ABS	take twelve times the kids version high roller my life told	Page was member follow a member opened fire worshipers at	follow say loved incredible feeling find an old jacket	discusses explain support his speech that lead
context-EXT	But for High Roller CEO Matt Armbruster, these <u>childhood</u> memories are priceless.	Police in Oak Creek had no contact with Page prior to Sunday's <u>shooting</u> <u>shipwreck salvage company</u> 1715 Fleet Queens Jewels, LLC, led his crew of three on an expedition off the <u>Florida coast</u> to in <u>New York</u> , one of the protesters at the <u>Occupy Wall Street gathering</u> jumped up onto a concrete benches on the north side of <u>Zuccotti Park</u> ...

We can safely conclude that the incorporation of *context* improves the performance of caption generation process. METEOR, which correlates highly with human evaluation, confirms that the proposed *context*-driven *extractive* strategy (*context*-EXT) generates the best captions. The best performance in terms of TER, among strategies generating captions of similar length, is also achieved when *context* is incorporated in the process, i.e., *context*-EXT and *phrase*-ABS(*context*-AIA).

Since our caption generation framework *context*-EXT *extracts* sentences from articles to be used as captions, the generated captions commonly refer to the news stories. In some cases, the ground truth caption is simply a named entity (a proper noun describing a person, location or organization) while the generated caption is a sentence describing the news story about the given named entity. Table 4.10 shows a few examples of such images. In such cases, the generated caption is relevant to the news image in its given *context*, even though it is not exactly the same as its ground truth caption.

Table 4.11 presents a comparative analysis of captions generated by our system (*context*-EXT) against various other systems like *phrase*-ABS and *baseline*-EXT(MBRM). It is apparent that the *context*-EXT has the ability to *extract* the best sentence to describe the image in the *context* of its associated news story or article. Main weakness of the phrase-abstractive technique (*phrase*-ABS) proposed by Lapata et al.[33], is the generation of incomprehensible sentences. This technique tries to build a sentence from scratch but ends up generating grammatically incorrect word sequences. *Context*-sensitive caption generation on the basis of joint image-words probability estimated by multiple Bernoulli relevance model (*baseline*-EXT(MBRM)) extracts descriptive sentences from the accompanying articles but the sentences selected by *context*-EXT are more suitable descriptions for images in the *context* of their articles. It is proven by the observation that the captions selected by *context*-EXT system overlap with the ground truth captions more than the captions generated by any other system.

4.3.3 Study of *Semantic Gap* in News Images

As shown in Figure 4.3, news image captions describe both the *contents* and the *context* of images. On the other hand, Figure 4.1 and 4.2 clearly demonstrate that the *context* is largely removed from ‘artificial’ image descriptions of datasets like IAPR TC-12 and MSCOCO.

LeCun et al. proposed a convolutional neural network (CNN) for automatically learning image representation vectors for the purpose of hand-written digit recognition[65]. In recent past, variations of this framework have been trained over the ImageNet database in which each image is labeled according to the object it shows[23]. Image representation vectors learned from such ImageNet-trained CNN perform extremely well for the systems dealing with the tasks of object recognition and image annotation. This technique seems to bridge the gap between image features and the semantic concepts as high as the labels of the ImageNet dataset, i.e., the names of everyday objects like ‘car’, ‘chair’, etc.



(a) Traders work on New York Stock Exchange floor. (**monitor, CRT screen**)



(b) Steve Jobs institutionalized his vision at Apple.(**stage, mike**)



(c) BestBuy CEO Brian Dunn resigned amid an investigation into his (**suit of clothes**)

Figure 4.8: TIME image-caption pairs; CNN-assigned ImageNet labels are written in bold face.

Image features extracted from ImageNet-trained CNN, for images of datasets like MSCOCO, IAPR TC-12 and ESP, are effective for predicting annotations for such images. The predicted annotations closely match their ground truth annotations. Note that the ground truth annotations

for images of these datasets are basically the names of common objects shown in the images like ‘car’, ‘chair’, ‘box’, ‘man’, etc. Hence, these ground truth annotations resemble the ImageNet class labels. The last layer of popular CNN frameworks corresponds to 1000 labels of ImageNet. The highest weighted ImageNet labels for sample TIME images are indicated in bold font in Figure 4.8. Though meaningful in terms of the visual contents, these labels have no correlation with the ground truth image captions because of the additional ambiguity in textual descriptions caused by the *context* of images. Hence, the image features extracted from ImageNet-trained CNN for images of the TIME dataset fail to perform well when incorporated in image annotation systems.

Table 4.12: Visual features comparison for image annotation (‘conv5’: last convolutional layer of CNN, ‘fc7’: last fully connected layer of CNN, ‘grid’: grid-based visual features)

	Visual features	Mean Precision	Mean Recall	N⁺
IAPR TC-12	Deep rep.[57]	42	29	252
	RandForest[35]	44	31	253
	2D-BoW[74]	24	26	145
	semanticBoW[75]	41	33	—
	<i>context-AIA(grid)</i> [111]	55	20	254
	<i>context-AIA(fc7)</i>	63	27	259
ESP	Deep rep.[57]	38	22	228
	RandForest[35]	45	24	239
	<i>context-AIA(grid)</i> [111]	45	19	246
	<i>context-AIA(fc7)</i>	61	21	245
Flickr30K	<i>context-AIA(grid)</i>	13	7	98
	<i>context-AIA(fc7)</i>	35	18	179
MSCOCO	<i>context-AIA(grid)</i>	11	4	93
	<i>context-AIA(fc7)</i>	49	19	235
TIME	<i>context-AIA(grid)</i>	44	21	758
	<i>context-AIA(conv5)</i>	42	21	748

Table 4.12 shows the results for our *context*-sensitive annotation model (*context-AIA*) with various visual features for standard image annotation datasets and the TIME dataset. We only employed scene-based *context* estimation strategy as it is the only common available *context* source for all of

these datasets. CNN features substantially improve the performance for standard image annotation datasets. Our *context*-AIA model even outperforms more recently proposed annotation models involving random forests generation[35], deep hierarchical learning[57] and modifications of bag-of-words representations[74, 75] for IAPR TC-12 and ESP datasets with ‘fc7’ features (features extracted from the last fully-connected layer of the CNN).

On the other hand, CNN features do not perform well for annotation of images of the TIME dataset. Mean precision of annotation system with ‘fc7’ features for images of the TIME dataset, was less than 10%. We experimented with ‘conv5’ features (features extracted from the last convolutional layer) for images of the TIME dataset. The performance of the annotation system was slightly worse than simple grid-based features. These results have important implications for recently proposed CNN-RNN based automatic image description generation systems like NeuralTalk[54].



Figure 4.9: News image from the TIME dataset; Ground-truth caption is “*Facebook CEO Mark Zuckerberg*”, NeuralTalk-generated caption is “*A man is sitting on the rock*”, Caption generated by *context*-EXT is “*Facebook was originally not created to be a company, Zuckerberg wrote in Facebooks prospectus.*”

The state-of-the-art CNN-RNN based caption generation systems like NeuralTalk[54] employ ImageNet-trained CNN as their initial processing module. The results in Table 4.12 indicate the inappropriateness of such CNN for generating image representations of news images for predicting

their textual annotations. These results explain the ineffectiveness of systems like NeuralTalk for generating captions of news images. When we employed NeuralTalk to generate the captions for the news image of the TIME dataset, the generated captions were somewhat relevant to the visual contents of images but had little correlation with the real world ground truth captions. Figure 4.9 shows a news image along with its real world caption, the caption generated by NeuralTalk, and the caption produced by our framework (*context*-EXT). It is clear that NeuralTalk-generated caption cannot replace the real world caption, while the caption produced by *context*-EXT is relevant to the image given its specific context.

TIME dataset presents a classic challenge of *transfer learning*. *Transfer learning* deals with the cases where training and test dataset (ImageNet and TIME datasets, respectively), are different from each other[91]. ImageNet-trained CNN have been fine-tuned for Flickr Style dataset by adding a new last layer whose entries correspond to the labels of the new dataset. When such fine-tuning is attempted for the TIME dataset, the network loss keeps growing instead of decreasing. Textual labels of the TIME dataset are characteristically too different from the labels of the ImageNet database, to properly fine-tune the pre-trained CNN. No large enough collection of news image-caption pairs is available to effectively train a CNN from scratch. Though, our dataset is a step in the direction of collection of such dataset.

4.4 Semantic Network of Named Entities

The main objective of our work is to develop meaningful understanding of the relations between images and text through the exploration of contextual cues from all available sources of semantic contextual information. Earlier, we discussed our ideas to deal with a major image-text relations problem, i.e., automatic image annotation and caption generation. In case of news images, ‘text’ is not limited to the sentences directly associated with images (i.e., captions). Instead, there is vast

amount of text available as news articles. This text has indirect but important relations with news images. News dataset is not the only example of this scenario. Various datasets collected from internet have free-flowing text implicitly associated with images. Such text can be in the form of a blog, a Wikipedia article, or general text available on webpages. Any system that aims at developing deep understanding of image-text relations for such datasets, can not afford to ignore the invaluable semantic information encoded in long sequences of text such as articles or blogs.

Establishing rigorous links between images and long sequences of text like news articles, may not be directly possible. The information encoded in articles needs to be distilled to a form that can be directly associated with images. Articles contain words from a very large vocabulary set but some of these words form special linguistic features that carry special meaning. As describes at the start of this chapter, named entities (names of people, places and organizations) are an example of such linguistic features. A study indicated that the named entities constitute the most common query terms used for searching through blog databases[82]. The results of this study are a clear proof of the importance of the named entities in large free-flowing text such as news articles or blogs.

Quite a few computer vision problems deal with the visual aspect of named entities, e.g., facial recognition [5, 89, 110, 46] for ‘person’ entities, landmark identification [16, 17, 1, 44] for ‘place’ entities, and logo recognition [100, 14, 29, 55, 101] for ‘organization’ entities. Given the limited amount of semantic information encoded in the visual contents of images, such systems mainly deal with identifying the visual representation of the named entities. For news or blogs datasets, there is huge potential for mining semantic information regarding these entities from the text available in news articles or blogs. Semantic information regarding such entities, extracted from the text, can further enrich the image-entity relations discovered by the above mentioned systems.

We devised a framework to automatically build a semantic network of named entities. Our framework extracts the semantic concepts and finds their association with the named entities. Named

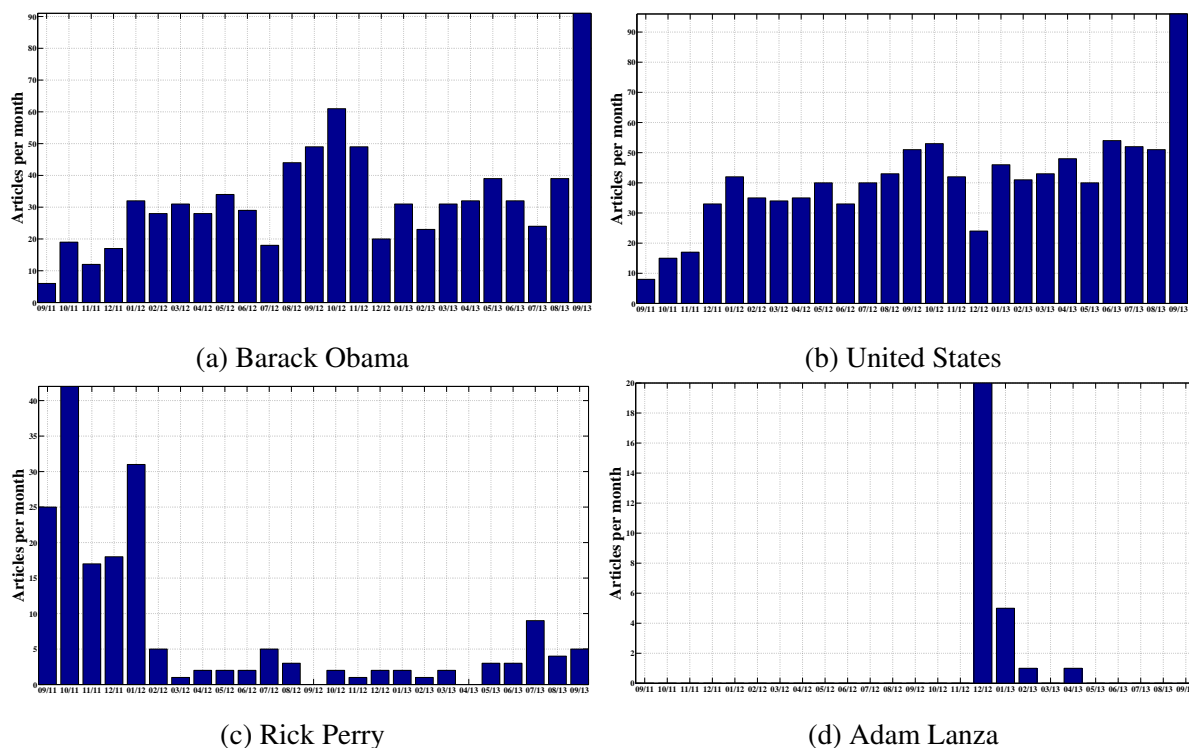


Figure 4.10: Distributions of named entities over time; each bar represents frequency of a named entity during one month.

entities linked to the same semantic concepts or topics are considered connected to each other. In the case of news datasets, the semantic concepts have special meaning as they may correspond to real world events or types/classes of events such as war, economy, sports, etc. Our systems has vast potential to be combined with image-entity relation extraction systems mentioned earlier. Such a unified system will cover both the visual and the semantic understanding of a very important linguistic feature, i.e., the names of people, places and organizations.

Figure 4.10 shows the distribution of four different named entities over a long period of time in news articles published in the TIME magazine. Entities such as the ‘United States’ and ‘Barack Obama’ occur frequently throughout the time span of more than three years (from April 2010 to October 2013). Other named entities such as ‘Adam Lanza’ and ‘Rick Perry’ occur frequently only

during small portions of the analyzed time period. In general, distributions of named entities vary greatly over time, e.g., distribution of ‘Adam Lanza’ has a sharp peak for a very short period of time corresponding to the event of Sandy Hook elementary school shooting incident in Newtown, Connecticut, USA. Such varying distribution provides a hint to the evolutionary nature of named entities mentioned in news articles. It is necessary for the system to be able to keep up with this evolution while tracking semantic relation of named entities.

4.4.1 *Sparse Structured Modeling for Named Entities Relations Extraction*

In keeping with our efforts to build frameworks that can automatically extract and identify semantic contextual relations of data items, we present a model to estimate the semantic background of named entities automatically through the assessment of words used with those named entities. We argue that the words used in the articles define a semantic background for the named entities mentioned in those articles. If two named entities are mentioned in the same type of semantic background, they must have a connection with each other. In this case, the common semantic background can define the *context* of the relation between the two entities.

We model named entity’s occurrences via sparse structured logistic regression. Such modeling ensures the identification of words that can strongly predict the presence or the absence of named entities while filtering out less-relevant words as noise. We argue that the words with strong correlation to prediction of named entity’s occurrence, define its *context* or the semantic background. We induce group structure in our model such that each group of predictors/words define a certain semantic topic. Thus, the structured sparse logistic regression model can identify the semantic topics that define the *context* of named entities as they have strong correlation with occurrence of named entities. Our framework defines and quantifies semantic relations between named entities through the identification of common semantic topics defining their *context*. We refer to our

framework as ‘NELasso’.

In compliance with the notations introduced in Section 4.1, we introduce the problem of extracting semantic relations between named entities. The system is given a news articles collection \mathcal{D} and the set of named entities \mathcal{E} mentioned in these articles. An article $d \in \mathcal{D}$ can contain zero or more named entities $e_i \in \mathcal{E}$. The vocabulary set \mathcal{W} that contains all words used in articles collection \mathcal{D} , is split into $K > 1$ groups such that each group defined a semantic topic. The system seeks to establish relations $r_{ij} = rel(e_i, e_j)$ between named entities e_i and e_j ($i \neq j$) based on vocabulary word groups that can strongly predict occurrence of these entities in news articles. Each relation is characterized by its *type*, $type(r_{ij})$, and its *strength*, $str(r_{ij})$, where the *type* qualifies the *context* of the relation and the *strength* quantifies it. Intuitively, a relation r_{ij} is likely to exist in \mathcal{D} when both entities e_i and e_j are mentioned in the same *context* (e.g., event) in \mathcal{D} . We formulate this intuition to discover and characterize relations between named entities.

4.4.1.1 Semantic topics

We devised multiple frameworks for grouping words used in articles in such a way that each word group defines a semantic concept or a topic or a news event. The following are the possible sources that can provide evidence for a semantic concept or news event.

4.4.1.1.1 Co-occurrence-based word groups

The co-occurrence of words in articles can be used to form groups of related words. Typically, each article discusses a specific topic or news story. The presence of a word in an article indicates its relationship with the topic or story of the article. Two articles that contain similar words are likely discussing the same topic or news story. Thus, co-occurrence of vocabulary words in the

same set of articles is an important clue for forming word groups. Our framework clusters words based on their occurrence in different articles $d \in \mathcal{D}$ such that words occurring in the same subset of articles are put together in one group. Each word group indicate a semantic topic or an event that is common between all articles of the corresponding subset.

To find co-occurrence-based word groups, we represent each word w_j in the vocabulary set by a vector \mathbf{v}_j of length M where M is the number of articles in the collection. The i_{th} element v_{ji} of this vector indicates presence (1) or absence (0) of the word w_j in the i_{th} article. Our system employs agglomerative hierarchical clustering of these vectors to find groups of related words. The cosine similarity is adopted for comparing vectors; the single-link merge operator is used; and a constraint is imposed to restrict cluster size to τ or less. Iteratively, any cluster larger than τ is further divided. This procedure results in a finite number of non-overlapping subsets \mathcal{W}_k ($k = 1, \dots, K$) of the vocabulary set \mathcal{W} such that $\forall k, |\mathcal{W}_k| \leq \tau$ and $\forall l, t, \mathcal{W}_l \cap \mathcal{W}_t = \emptyset$. The threshold τ determines the maximum allowed size of the word groups, and hence the number of word groups formed.

4.4.1.1.2 *Keyword-based word groups*

News websites often assign one or more keywords to each article which characterize its topical *context* and help the reader navigate to other articles discussing the same topic. Words appearing in articles having a certain keyword are obviously indicative of the topic or news event represented by that keyword. Thus, keywords can aid the process of identifying groups of words associated with a topic.

We estimate the importance of each vocabulary word in identifying a particular topic represented by a specific keyword. This scenario is similar to the term-to-topic relatedness concept introduced in [115, 53]. Relationships between words, i.e., term-to-term relationships are commonly used in

many natural language processing tasks. However, relationships between words and topics, i.e., term-to-topic relationships are more useful in cases where *context*/topic is already known. In our setting, topic or *context* is specified in the form of keywords assigned to each news article.

The relatedness of a word in the vocabulary set with a *context* defined by a keyword can be quantified by its discriminative term weight (*dtw*). The *dtw* for vocabulary word w_j given *context*/keyword key is defined as

$$dtw(w_j, key) = \frac{p(w_j|key)}{p(w_j|key')} \quad (4.39)$$

$p(w_j|key)$ is the probability of word w_j in articles associated with keyword key while $p(w_j|key')$ is the probability of word w_j in all other articles[115, 53]. To estimate these probabilities, we assume a document model in which each word follows the Bernoulli distribution, i.e., the word either occurs or does not occur in articles of a given keyword. Each word is associated to the keyword for which it has the highest *dtw*. Let key_j denotes the keyword to which word w_j has been assigned, then

$$key_j = \arg \max_{key} dtw(w_j, key) \quad (4.40)$$

Thus, each vocabulary word is assigned to one keyword. \mathcal{W}_k represents a subset of the vocabulary set \mathcal{W} consisting of all words assigned to the keyword key_k . These words are indicative of the semantic topic defined by the corresponding keyword. The resulting word-groups are non-overlapping, i.e., $\forall l, t \mathcal{W}_l \cap \mathcal{W}_t = \emptyset$

Oftentimes, the distribution of articles among keywords can be extremely uneven. Some keywords, such as ‘World’ (in TIME dataset), are too general and are assigned to a large number of articles covering many different topics. Thus, the word groups for such keywords are very large. To address this issue, we further divide word groups \mathcal{W}_k with $|\mathcal{W}_k| > \tau$ into smaller word groups

using co-occurrence pattern of words, as described in the previous paragraph. For example, the word group corresponding to the keyword ‘World’ may now be divided into subgroups ‘World1’, ‘World2’, etc.; each subgroup corresponding to one news story covered by articles of keyword ‘World’. In this process, the threshold τ determines the maximum allowed size of the word groups and affects the number of word groups formed, as it does for co-occurrence based word groups.

4.4.1.1.3 Topic-based word groups

Topic modeling is a powerful tool for document collection understanding[10]. Through latent Dirichlet allocation (LDA) based modeling, documents are treated as mixtures of topics while each topic is defined as distribution over words. Section 4.2.1.1 describes the step involved in such modeling. Since our system requires identification of word groups belonging to certain topics, we employ a similar modeling scheme where topics are defined as the probability distribution over all words of the vocabulary, i.e., $P(w_j|C_k, \xi)$ (ξ is a fixed quantity to be estimated by the topic modeling process). We take the set of underlying topics of the article collection as the basis for word group formation. The system uses a threshold ε on the value of $p(w_j|C_k, \xi)$ to decide whether or not the word w_j belongs to the topic C_k . Thus, there are as many word groups as the number of topics (K). The group corresponding to topic C_k contains all the words with reasonably high conditional probability given C_k . In general, this method forms overlapping word groups. Each word group \mathcal{W}_l is a subset of vocabulary set \mathcal{W} such that $\mathcal{W}_l \cap \mathcal{W}_t \neq \emptyset$. For this method, our system repeats words appearing in multiple word groups in the vector \mathbf{d}_m for the m_{th} article.

4.4.1.2 Sparse Structured Modeling

The occurrence of a named entity in news articles depends on the *context* (topic, event, story, etc.) of the articles in which it is mentioned, and the *context* is specified by the words used in those

articles. We use this idea to model the occurrence of each named entity as a classification problem, where the words appearing in articles serve as predictors and the occurrence of the named entity in the articles as the target. A separate model is learned for each named entity in \mathcal{E} .

Not every word plays a significant role in predicting every entity. We adopt a sparsity inducing approach by using ℓ_1 -norm of coefficients as a penalty to the standard classification objective function. Our framework forms word-groups such that words in one group are semantically similar. We also impose a penalty on all the coefficients of words from each group. This penalty is the ℓ_2 -norm of the coefficients of each word group. It tries to eliminate entire groups of words from the model, further enhancing sparsity and interpretability of the model, especially when groups carry contextual semantics. The sparse group lasso logistic regression model for the named entity e is given as

$$\min_{\mathbf{x}} \left[\sum_{m=1}^{M_e} \ln(1 + \exp(-y_m(\mathbf{x}^T \mathbf{d}_m + c))) + \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{k=1}^K \phi_k \|\mathbf{x}_k\|_2 \right] \quad (4.41)$$

Here, $y_m \in \{-1, +1\}$ indicates whether m_{th} article mentions the entity ($y_m = +1$) or not ($y_m = -1$). M_e is the number of articles used in training. In practice, we prefer to have articles that mention and do not mention the entity in almost equal proportions in the training set; thus, the training set for $e \in E$ includes all articles that mention e and the same number of randomly picked articles that do not mention e . Therefore, $M_e \leq M$ in general. The vector $\mathbf{d}_m \in \mathbb{R}^N$ represents the m_{th} article in bag-of-words format. The vector $\mathbf{x} \in \mathbb{R}^N$ contains the learned coefficients corresponding to the words in \mathbf{d}_m .

The model assumes a group structure among words such that the coefficient vector \mathbf{x} consists of K non-overlapping groups of coefficients \mathbf{x}_k . The term ϕ_k assigns an additional weight/penalty to the k_{th} group of coefficients. These terms can be selected empirically, but in most cases in practice (including our experiments), they can be set to one. There are two regularization parameters or terms in the ℓ_1/ℓ_2 regularized logistic regression model. The first term λ_1 rewards the selection

of fewer words, while the second term λ_2 enforces sparsity on the group structure of the words – it rewards selection of as few groups as possible from the available groups of words. The sparse group lasso model can be solved efficiently by the implementation provided in the SLEP package⁵. This implementation also finds the optimal values of the regularization parameters automatically.

It is interesting to note that, out of three methods of feature group formation discussed in Section 4.4.1.1, one forms overlapping groups, i.e., the features are repeated among different groups. We form the data matrix \mathbf{A} for this case by repeating the shared features. Thus columns of the data matrix are not strictly independent of each other. We have noticed that model still estimates reasonably good relationships between the response and the input vectors.

A sparse group-structured model, i.e., the vector \mathbf{x} containing the coefficients corresponding to the words in the vocabulary set, is estimated for each named entity in \mathcal{E} . This information, together with how these coefficients exist across groups, is used to establish relations among named entities. With this information, we also define the *type* as well as the *strength* of each relation.

A word provides positive evidence for a named entity e if the value of the corresponding coefficient in the entity’s prediction model is greater than zero. The evidence provided by words in the k_{th} group for entity e , denoted by t_k^e , can be estimated by summing up entries \mathbf{x}_n of the coefficient vector \mathbf{x} such that this n_{th} word/coefficient belongs to group k and $\mathbf{x}_n > 0$. Relation between entities e_i and e_j are formed based on the common word groups or semantic topics that provide positive evidence for both entities. The strength of this evidence will be used to define the *strength* of relation between the two entities. The *type* of such relation will be based on the semantic topic presented by the common word group.

We say that this evidence is significant when it is greater than a threshold, i.e., $t_k^e \geq \gamma$ where $\gamma \geq 0$

⁵<http://www.public.asu.edu/~jye02/Software/SLEP/>

is a selection threshold. The value of γ decides the amount of positive evidence a group of words needs to provide for a named entity for it to be considered as a contender for establishment of semantic relations.

Consider a relation r_{ij} between entities e_i and e_j . The *types* of this relation are given by the groups of words that provide significant positive evidence for both entities e_i and e_j . For example, if the group defined by the keyword ‘Election’ provides significant positive evidence for named entities ‘Mitt Romney’ and ‘Paul Ryan’ then the *type* of the relation between these named entities is ‘Election’. In general, one or more *types* can characterize a relation. If no groups provide significant positive evidence for both of the entities, then no relation exists between them. The *strength* of relation r_{ij} of type k is defined as $str_k(r_{ij}) = t_k^{e_i} \times t_k^{e_j}$.

Definition 1 (*Relation r_{ij}*) A relation r_{ij} of type k exists between entities e_i and e_j in \mathcal{D} when both $t_k^{e_i}$ and $t_k^{e_j}$ are greater than the selection threshold $\gamma \geq 0$. Here, t_k^e is the sum of the positive coefficients in the k th group in the sparse group logistic regression model for the entity e .

The strength of r_{ij} is defined as $str_k(r_{ij}) = t_k^{e_i} \times t_k^{e_j}$.

Semantic relations between named entities are not dependent on their association with individual words but with groups of similar words. Each group is the lexicon for one particular *context* or topic. This ensures that relations between named entities are not ignored because of the use of different individual words. They are estimated based on whether or not both named entities relate to the lexicon of the same *context*, thus enabling the system to identify more complex relations.

Time dependence is induced by estimating this regression model for each entity over articles of a certain time period. Any identified relations between named entities are valid for the given time period only. Relations evolve when regression model for entities are estimated over the set of

articles from the next time period. Given the evolutionary nature of news material, it is important for relation extraction system to be able to update relations based on time.

4.4.2 System Output

The system is given a news articles collection along with its metadata, e.g., article publication date and keywords. We used the TIME and the BBC datasets (Section 4.1) to generate the output of our system. The text of news articles is processed to form a vocabulary set and the named entities mentioned in the articles are identified. As explained earlier, TreeTagger is used to tokenize, lemmatize and part-of-speech tag news articles' text. Frequently occurring nouns, verbs, and adjectives are retained to form the vocabulary set \mathcal{W} . Each article is represented as a vector \mathbf{d} of length N where $N = |\mathcal{W}|$. Each element of \mathbf{d} records the number of times the corresponding word set \mathcal{W} occurs in an article. This is standard bag-of-words representation for text documents. There are many techniques for identifying named entities reliably from text documents. Our system uses the Stanford named entity recognizer (NER)⁶ trained over MUC named entity corpora that identifies 7 different classes of entities, i.e., Person, Organization, Location, Time, Percent, Money, and Date. The system retains only named entities of types Person, Organization, and Location as these are the most interesting and important entities mentioned in news articles.

4.4.2.1 Network of Named Entities

The relations among named entities in a given time period can be presented visually as a semantic network. Figures 4.11 and 4.12 are two sample semantic networks generated by our system for two different time slots of the TIME dataset. An edge between two entities indicates a relation

⁶<http://nlp.stanford.edu/software/CRF-NER.shtml>

between those entities. The thickness of an edge indicates the *strength* of the strongest *type* of relation between the entities. ‘Gaza’, ‘Hamas’, ‘West Bank’, ‘Tel Aviv’, ‘Jerusalem’, and ‘Israel’ are connected to each other with thick edges in Figure 4.11. This network corresponds to the time of the ‘Operation Pillar of Defense’ which involved these entities. ‘Edward Snowden’, ‘NSA’, ‘Ecuador’, and ‘Hong Kong’ are connected to each other in Figure 4.12. This network corresponds to the time when the NSA leaks story broke out. Such networks generated by our system provide intuitive understanding of news stories and named entities discussed in a given time period.

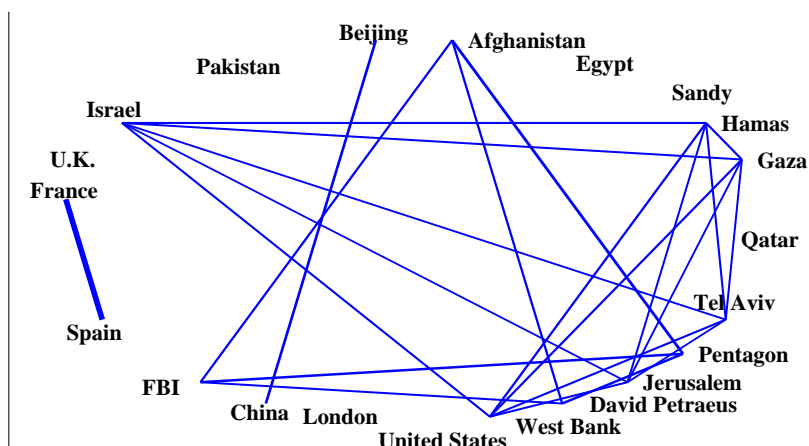


Figure 4.11: Semantic network of named entities for Nov-Dec 2012 (TIME dataset)

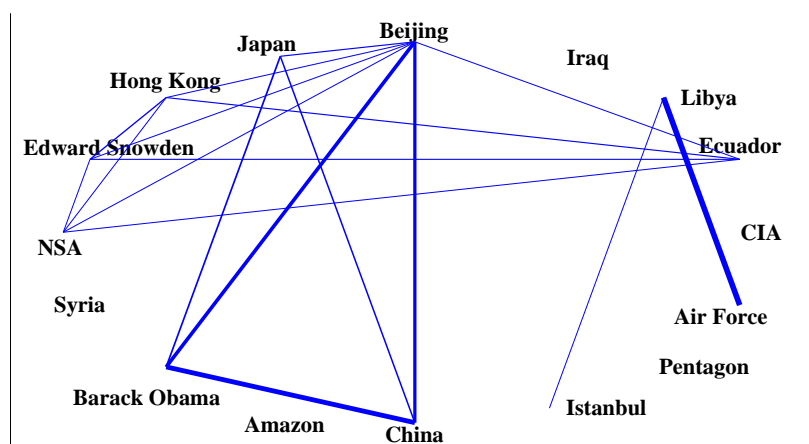


Figure 4.12: Semantic network of named entities for Jun-Jul 2013 (TIME dataset)

4.4.2.2 News Events

In the semantic networks built by our system, we can identify cliques of related named entities. A clique in a network is a group of named entities in which every named entity is related to every other named entity in the group and all relations are of the same *type*. These cliques typically correspond to major events in the news articles' dataset, and provide a summary at a glance of named entities involved in the events. Table 4.13 gives some cliques identified by our system along with the time period in which they occur and their *type*. For keyword-based word groups, the keyword provides a label for the relation *type*. For co-occurrence- and topic-based word groups, relation *type* has been indicated by a few top words of the word group responsible for the connection among the named entities. In any case, the relation *type* points to the news story in which the named entities of the corresponding clique play important roles.

Table 4.13: Example cliques discovered by our system (TIME dataset); each clique corresponds to a distinct news event indicated by the *type* of the relation

Named Entities	Time Period	Relation <i>type</i>		
		Keyword	Co-occurrence	Topic
Colorado, James Holmes, Aurora	Jul-Aug,2012	Crime	Aurora, Shooting, Theater	Kill, Shooting, Colorado
South Korea, Pyongyang, North Korea, Kim Jong II	Dec,2011-Jan,2012	North Korea	North, Korean, Korea Imperial, Successor	North, Korea, Leader Military , Dictator
Israel, Hamas, Tel Aviv, Gaza, Jerusalem, West Bank	Nov-Dec,2012	Israel	Gaza, Hamas, Radical Israel, Occupation	Israel, Palestinian, Fire Rocket, Refugee, Gaza

4.4.2.3 Dynamics of Relations

The output of our system makes it easy to understand the dynamics of relations among named entities over time. Figure 4.13 shows the variation of average *strength* of relations between all pairs of entities among a selected set of named entities over different time periods in TIME dataset. The

set of named entities (given in the figure’s caption) is selected such that relations between all pairs of these entities exist in all time periods. These graphs (one each for co-occurrence-, keyword-, and topic-based word groups) show that the average *strength* (blue line) varies greatly over time for the same set of relations. These graphs also show that the average WLM (Wikipedia link-based measure) across all pairs of entities (green line) remains constant over time as WLM is a static measure of relation strength derived from Wikipedia (see Section 4.4.3.1 for the details of WLM).

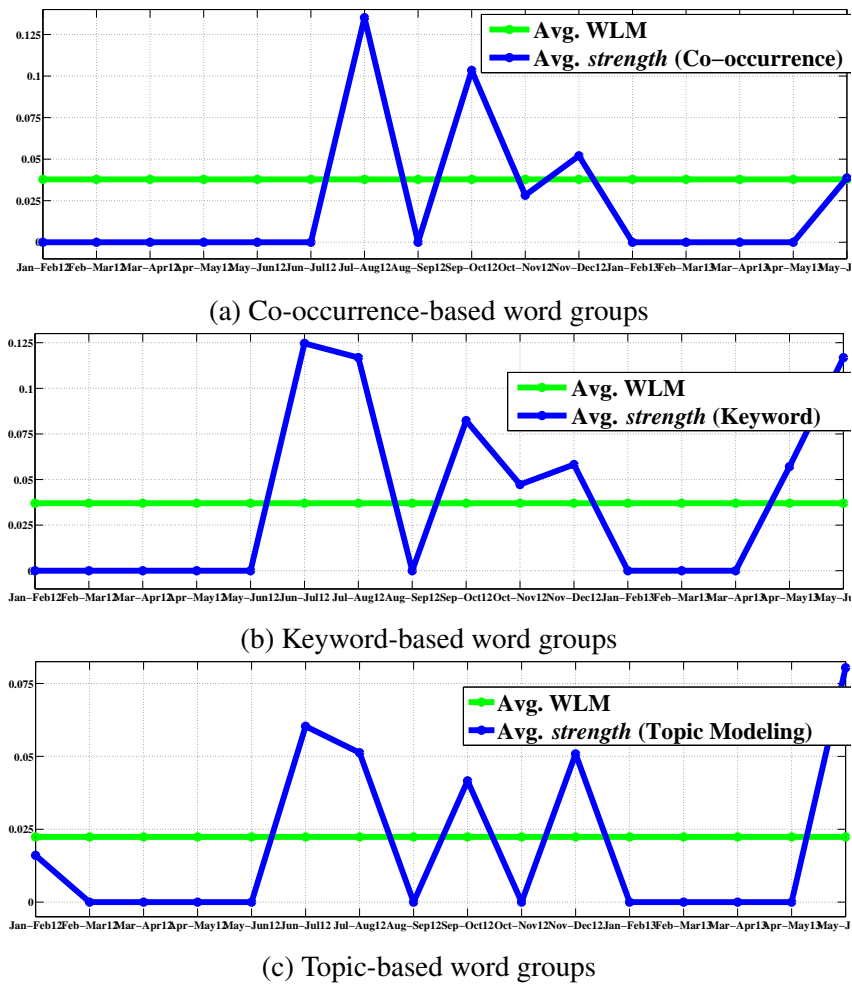


Figure 4.13: Variation of average *strength* and WLM of relations over time (TIME dataset). Relations among the following named entities exist in each time period: {Syria, Bashar Asad, Cairo, Damascus, Jerusalem, Hamas, Gaza, Israel Egypt, Benghazi, Hillary Clinton}; x-axis: Time period (months). y-axis: Mean WLM and relation *strength*

Table 4.14: Example of a named entity involved in different relations over time (TIME dataset). The *strength* of the relation using keyword based word groups is shown in parenthesis

Related Named Entities	Time Period	Associated News Story
Mitt Romney - South Carolina (0.04)	Jan-Feb, 2012	South Carolina Republican Primary: Jan 21, 2012
Mitt Romney - Florida (0.03)	Jan-Feb , 2012	Florida Republican Primary: Jan 31, 2012
Mitt Romney - Arizona (0.03)	Feb-Mar, 2012	Arizona Republican Primary: Feb 28, 2012
Mitt Romney - Ohio (0.05)	Feb-Mar, 2012	Ohio Republican Primary: Mar 06, 2012
Mitt Romney - Illinois (0.07)	Mar-Apr, 2012	Illinois Republican Primary: Mar 20, 2012
Mitt Romney - Paul Ryan (0.11)	Aug-Sep, 2012	Mitt Romney announced Paul Ryan as his running mate on August 11, 2012
Mitt Romney - Tampa (0.06)	Aug-Sep, 2012	Mitt Romney formally accepted Republican Party nomination on August 30, 2012 in Tampa, Florida.

Table 4.15: Sample relations with more than one relation *type* in one time period (TIME dataset)

Related Named Entities	Time Period	<i>type</i> #1	<i>type</i> #2
Spain - U.K.	Oct-Nov, 2012	BBC, Live, International, Set, European	Economics, Rise, Growth, Spending, Crisis
Mitt Romney - White House	Oct-Nov, 2012	President, Presidential , Debate, Obama, Romney	Election, Candidate, Vote, Poll, Race
Iran - Russia	Mar-Apr, 2013	Diplomat, Negotiation, Sanction, Suspension	Aggression, Ballistic, Firing, Hostile, Target
Turkey - Istanbul	Jun-Jul, 2013	Police, Protest, Street, Night, President	War, Syria, Rebel, Assad, Regime

The temporal variation in average relation *strength* can be linked to the popularity of news stories involving the selected named entities. The blue lines for all three types of word groups have distinct peaks in July 2012, corresponding to the news story about Damascus bombing involving named entities ‘Syria’, ‘Bashar Asad’, ‘Damascus’, etc. Peaks observed in September 2012, correspond to the Benghazi attack and its aftermath involving a discussion on entities such as ‘Damascus’, and ‘Hillary Clinton’. News story of Operation Pillar of Defense involving entities ‘Israel’, ‘Hamas’, ‘Gaza’, ‘Egypt’, etc., corresponds to the peaks observed in November 2012. Peak in May 2013 correspond to a rare interview of Bashar Asad involving entities ‘Syria’ and ‘Israel’. Our system successfully captures the evolutionary nature of named entities relations in news material.

Our system discovers various relations of ‘Mitt Romeny’ with other named entities over time (Table

4.14), correlating with the occurrence of certain news events. Thus, our system can track an entity over time, discovering its relations to specific events or news stories.

Table 4.15 shows pairs of named entities which are related to each other with more than one relation *types* at the same time. Each relation *type* hints at some news story involving both entities. Our system is flexible enough to deal with the complexity of news material based named entities' relations whereas static relation measures, e.g., WLM, fail to do so.

The outputs of NELasso highlight its suitability for news material understanding, and this is the primary purpose of this system. Previously proposed systems do not possess such a capability and have a different goal altogether, i.e., construction of databases of facts [124, 28, 102, 11, 130].

4.4.3 Evaluation for Named Entity Relations

The main output of our system is the semantic network of named entities for a time-period of interest. There are two quantitative characteristics of such networks; 1) average degree or connectivity of the network, and 2) average *strength* of the relations in the network. The degree or average connectivity is defined as

$$Connectivity = \frac{2 \times \Sigma}{\Upsilon} \quad (4.42)$$

where Σ and Υ are the numbers of the edges and the nodes in the network, respectively. Our system assigns *strength* to each discovered relation, i.e., $str_k(r_{ij})$ is the strength for relation r_{ij} of *type* k between entities e_i and e_j .

There are two major aspects for evaluation of the named entities relations extracted by our system.

1. The relation are verifiable through some independent source
2. The relations are useful for a search and retrieval engines.

We treat Wikipedia as the independent source of verification for relations between named entities. Wikipedia is an extensive and highly organized database of information that has been widely explored in the literature for named entities' relation extraction[109, 122]. Our Wikipedia-based evaluation measure is presented in Section 4.4.3.1.

Our second automatic method is aimed at evaluating the usefulness of extracted relations in terms of search and retrieval scenario. One example system is a news recommendation tool. Such a system should be able to search and retrieve news stories that may interest a user reading about a particular named entity in a certain *context*. We present our evaluation measure to judge the effectiveness of the extracted relations for retrieval systems in Section 4.4.3.2.

4.4.3.1 Wikipedia-based Evaluation

The first evaluation measure is designed to check if the relations found by our system can be verified through an independent source. Wikipedia is an extensive and highly organized database of information regarding named entities. It has been widely used for extracting and characterizing relations among named entities[109, 122]. Milne et al. proposed Wikipedia link-based measure (WLM) to quantify the relatedness between articles a_{wiki} and b_{wiki} based on their inward and outward links in Wikipedia[122]. Each outward link from a_{wiki} and b_{wiki} is assigned a weight given by

$$\omega(source \rightarrow target) = \log\left(\frac{|\mathcal{W}_{wiki}|}{|\mathcal{T}_{wiki}|}\right) \text{ if } source \in \mathcal{T}_{wiki}, 0 \text{ otherwise} \quad (4.43)$$

where *source* can be a_{wiki} or b_{wiki} and *target* can be other articles in Wikipedia. \mathcal{T}_{wiki} is the set of all articles that link to the *target* and \mathcal{W}_{wiki} is the set of all articles in Wikipedia. Two vectors are formed corresponding to the articles a_{wiki} and b_{wiki} such that the corresponding entries of these vectors contains the weights of common links of these articles. The relatedness between a_{wiki} and b_{wiki} is based on the angle between their respective vectors.

The relatedness between a_{wiki} and b_{wiki} based on inward links is estimated as

$$sr(a_{wiki}, b_{wiki}) = \frac{\log(\max(|\mathcal{A}_{wiki}|, |\mathcal{B}_{wiki}|)) - \log(|\mathcal{A}_{wiki} \cap \mathcal{B}_{wiki}|)}{\log(|\mathcal{W}_{wiki}|) - \log(\min(|\mathcal{A}_{wiki}|, |\mathcal{B}_{wiki}|))} \quad (4.44)$$

Here, \mathcal{A}_{wiki} and \mathcal{B}_{wiki} are sets of all articles that link to a_{wiki} and b_{wiki} , respectively. The final relatedness of a_{wiki} and b_{wiki} is the average relatedness based on the outward and the inward links.

In our setting, a_{wiki} and b_{wiki} correspond to the entities e_i and e_j . WLM of this pair of entities, denoted by $wlm(e_i, e_j)$, is calculated if this pair is determined to be related by our system. The higher the $wlm(e_i, e_j)$, the stronger is the verification of the relation $rel(e_i, e_j)$ identified by our system through a completely independent source, i.e., Wikipedia. Therefore, we report the mean of $wlm(e_i, e_j)$ for all pairs e_i and e_j identified to be related by our system in a given time period as an evaluation measure of the semantic network of entities built for that time period.

It is worth emphasizing that there are certain advantages to relation extraction though our system over Wikipedia based relation identification. Our system assigns a *type* to each relation and allows the relation between two entities to change its *strength* or *type* or both over time. For instance, entities ‘Mitt Romney’ and ‘Barack Obama’ are mentioned frequently in many time periods but relate to each other through relations of varying *type* and *strength* in different time periods. WLM is static over time and provides no clue about the type of relation between two entities. Hence, it is not meant not meant to judge the quality of relation *type* assigned by our system.

4.4.3.2 Retrieval-based Evaluation

This automatic evaluation measure judges the usefulness of the relations identified by our system in retrieval scenario. A relation r_{ij} between the entities e_i and e_j has a *type* and a *strength* based on some word group, say k . The word group characterizes the *context* of each relation. If a user is

reading a news article that mentions entity e_i in the same *context* as that of a relation $rel(e_i, e_j)$, she should be suggested to read other articles which match the same *context* and mention the entity e_j . Users reading about ‘Mitt Romney’ in articles related to the ‘Election’ should be suggested to read about ‘Paul Ryan’ in other articles of the same *context* (i.e., ‘Election’). The *context* of a relation needs to be quantified to implement such a recommendation system. We do this by proposing a statistical signature vector ψ^k of length N . Three forms of this vector are developed, one for each type of word group formation explored in our system. All entries of ψ^k are set to zeros except the i_{th} entry ψ_i^k if the corresponding $w_i \in \mathcal{W}_k$. In this case, ψ_i^k is equal to either

1. sum of *tfIdf* weights of w_i for all documents in the given time period for co-occurrence based group formation
2. $dtw(w_i, key_k)$ in the given time period for keyword-based group formation
3. $P(w_i|C_k, \xi)$ in the given time period for topic-based group formation.

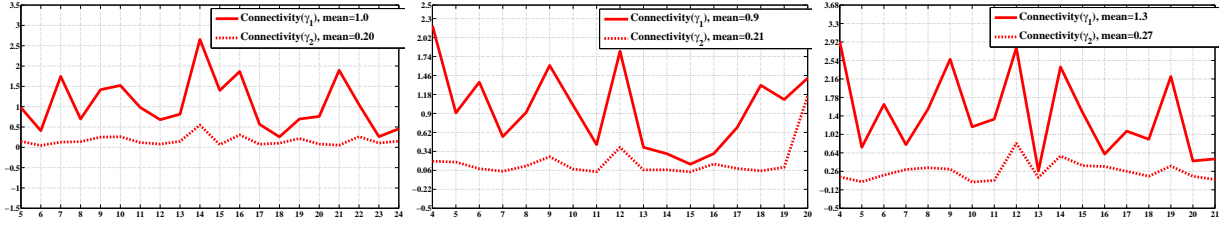
Using the above quantification of *contexts*, this evaluation method builds two lists of articles, l_i and l_j , from the given time period for each relation $rel(e_i, e_j)$ of *type* k identified in that time period. Articles in list l_i mention named entity e_i and match context ψ^k , whereas articles in the list l_j mention named entity e_j and match the *context* ψ^k . The match between a *context* and an article is determined by thresholding the cosine similarity between the vector ψ^k of the *context* and the bag-of-words representation vector of the article. We compute the percentage overlap between the two lists as an evaluation measure, called retrieval score, for the relation $rel(e_i, e_j)$ with *type* k . A higher overlap indicates that the topics of discussion regarding the two entities in the given relation are largely the same. Therefore, when a user who is reading about an entity e_i is recommended to read the articles about the entity e_j discussed in the same *context*, she will find the recommendations highly relevant.

Notice that this evaluation measure takes into account the *type* assigned to each identified relation based on a word group describing a *context* in the news articles.

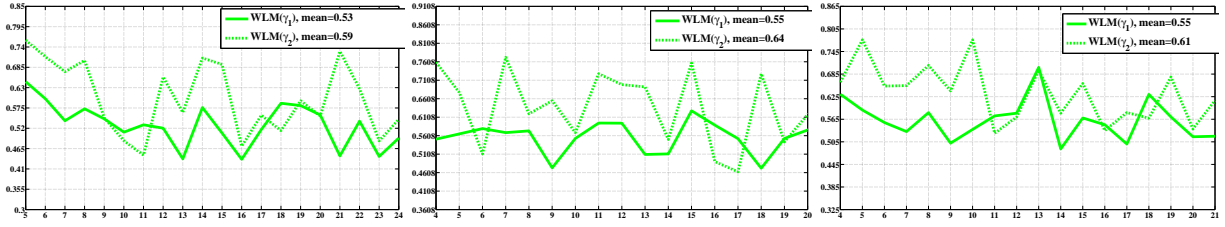
4.4.3.3 Automatic Evaluation Results

In this section, we discuss the results of automatic evaluation of our system. We start by discussing the impact of the parameter γ of the system. As discussed in Section 4.4.1.2, increasing the value of γ forces the system to pick named entities that have stronger evidence from word groups.

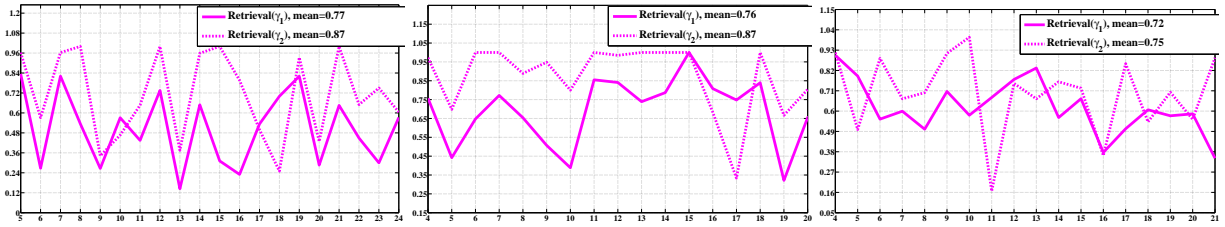
Figures 4.14 and 4.16 show the effect of γ on the evaluation measures for the TIME and the BBC datasets, respectively. In these figures, the x-axes represent the indices of one-month-long time intervals (TIME dataset) or indices of random subsets of the data (BBC dataset). One semantic network is built for each time interval or subset. The y-axes in these figures give the magnitude of various evaluation measures. The dotted lines are for a higher value of γ as compared to the solid lines. It is observed that the solid line is higher than the dotted line for mean connectivity, as increase in γ produces fewer relations. The dotted line is generally higher than the corresponding solid line for mean WLM, mean retrieval score, and mean *strength*, as increase in γ forces the system to pick relations with stronger evidence. These trends are consistent across both the datasets and all configurations of the system for every type of word group (co-occurrence, keyword, topic). Figures 4.15 and 4.17 depict the summary statistics for change (with increase in γ) in mean evaluation measures of semantic networks built for all time intervals for the TIME dataset and all subsets of the BBC dataset, respectively. Each boxplot shows the minimum, 25th percentile, 50th percentile (median), 75th percentile, and maximum of the change in the corresponding mean evaluation measure. A positive value indicates an increase in the mean evaluation measure. It is clear from these figures that increase in the mean WLM, retrieval score, and *strength* with increase in γ is the dominant trend as corresponding boxplots are above the zero-line.



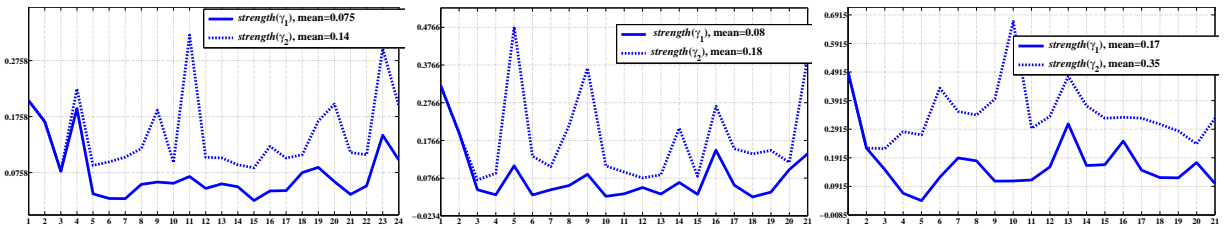
(a) Co-occ. word groups - Connectivity (b) Keyword word groups - Connectivity (c) Topic word groups - Connectivity



(d) Co-occ. word groups - WLM (e) Keyword word groups - WLM (f) Topic word groups - WLM



(g) Co-occ. word groups - Retrieval score (h) Keyword word groups - Retrieval score (i) Topic word groups - Retrieval score



(j) Co-occ. word groups - Strength (k) Keyword word groups - Strength (l) Topic word groups - Strength

Figure 4.14: Effect of threshold γ on evaluation measures on TIME dataset; x-axis: Time period (month), y-axis: Evaluation measure (connectivity, mean WLM, mean retrieval score, mean *strength*); Solid line: $\gamma = \gamma_1$, Dotted line: $\gamma = \gamma_2$ where $\gamma_1 < \gamma_2$; Mean of each curve given in legends.

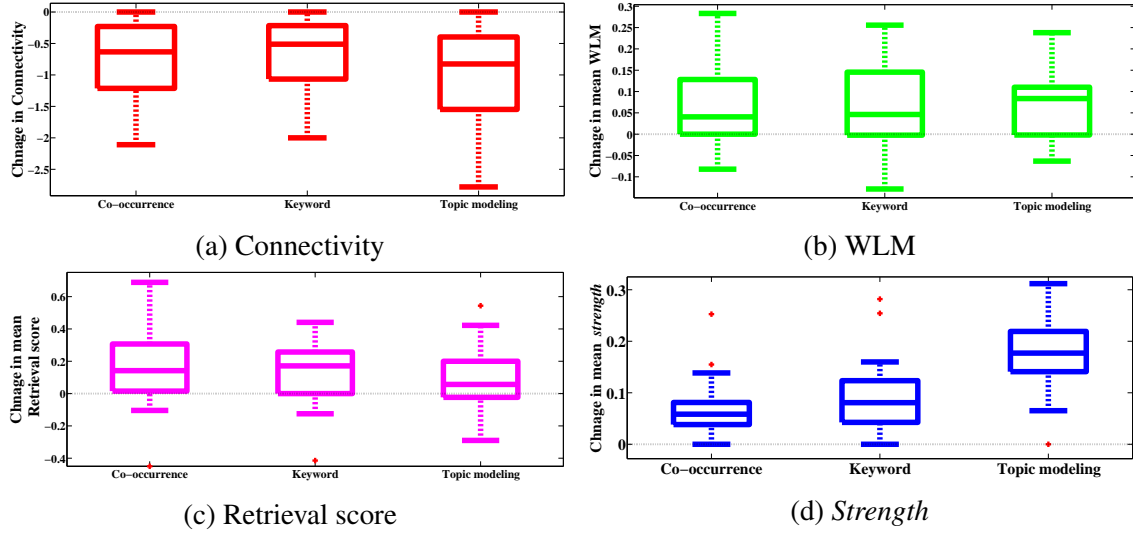
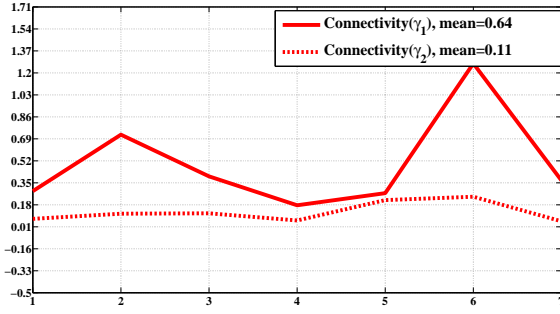


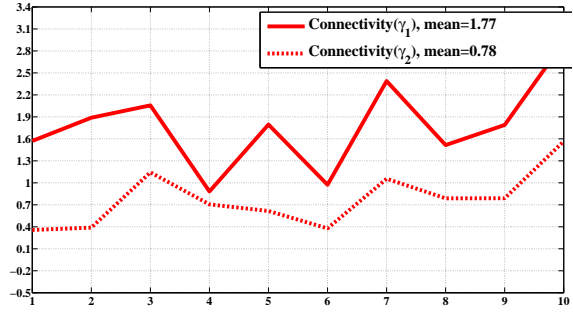
Figure 4.15: Summary statistics (boxplots) for connectivity, mean WLM, mean retrieval score, and mean *strength* at γ_2 minus that at γ_1 over all time periods (TIME dataset), $\gamma_2 > \gamma_1$

Note that the trend of change in mean evaluation measures for each semantic network is stronger in the TIME dataset (Figure 4.14) than that in the BBC dataset (Figure 4.16). This can be attributed to the fact that semantic networks on the TIME dataset are generated on articles published during one month. It ensures that many of the articles of one news story are available for processing together and results in more meaningful relations between named entities in the *context* of that story. On the other hand, the networks on the BBC dataset are generated on random subsets of the data with no guaranty of availability of significant information about one news story in one subset.

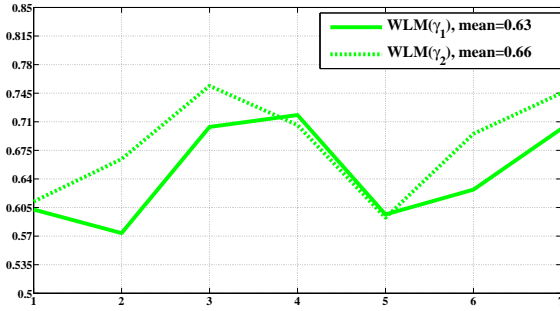
The parameters τ and K are discussed in Section 4.4.1.1. These parameters control the number of word groups formed by our word group formation methods. We observed that fewer word groups of larger sizes generate more relations. Hence, the connectivity of the system increases if fewer word groups are formed. However, the threshold γ affects both the connectivity and the quality of the generated networks regardless of the size and the count of the word groups. In other words, it controls both the number and the quality of the discovered relations. Thus, the choice of γ is more important than that of τ and K in our system.



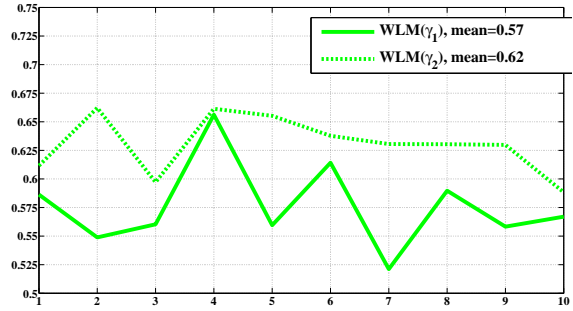
(a) Co-occurrence - Connectivity



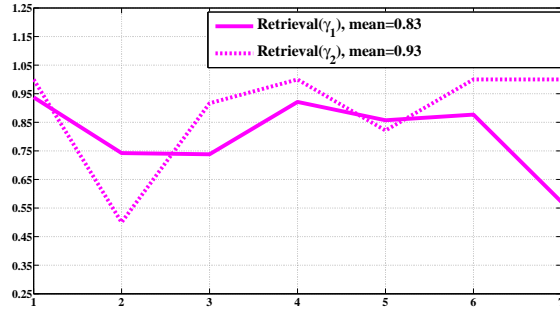
(b) Topic - Connectivity



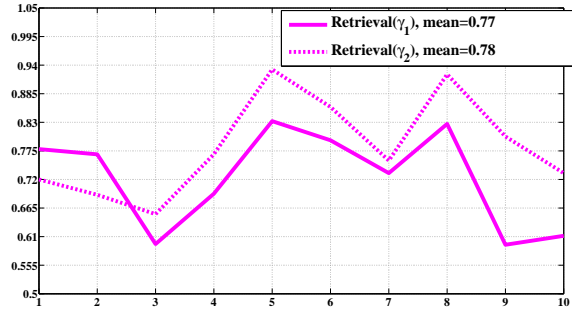
(c) Co-occurrence - WLM



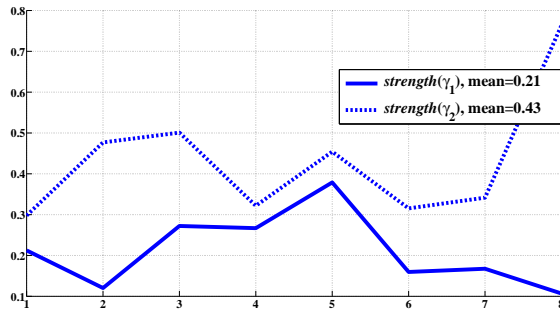
(d) Topic - WLM



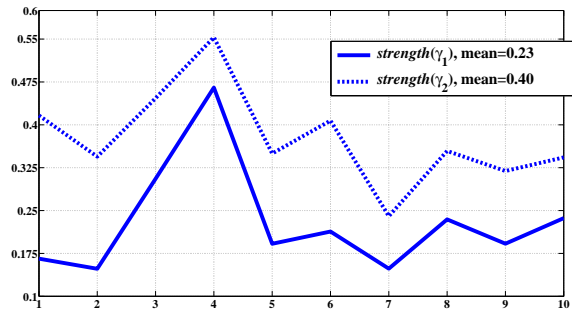
(e) Co-occurrence - Retrieval score



(f) Topic - Retrieval score



(g) Co-occurrence - strength



(h) Topic - strength

Figure 4.16: Effect of threshold γ on evaluation measures on BBC dataset; x-axis: Dataset sample, y-axis: Evaluation measure (connectivity, mean WLM, mean retrieval score, or mean *strength*); Solid line: $\gamma = \gamma_1$, Dotted line: $\gamma = \gamma_2$ where $\gamma_1 < \gamma_2$; Mean of each curve given in legends.

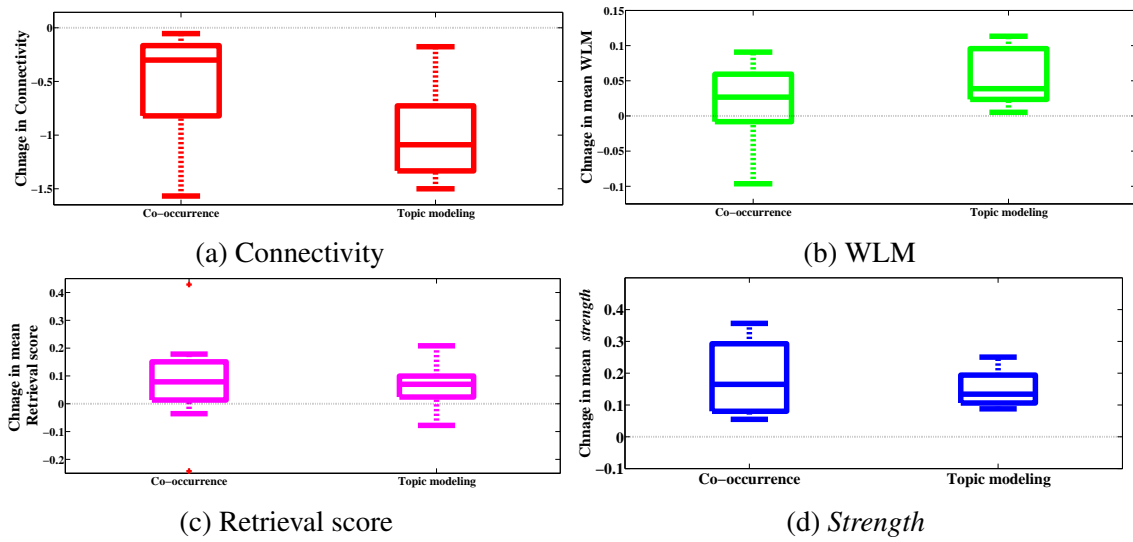


Figure 4.17: Summary statistics (boxplots) for connectivity, mean WLM, mean retrieval score, and mean *strength* at γ_2 minus that at γ_1 over all samples (BBC dataset), where $\gamma_2 > \gamma_1$

4.4.3.4 Effects of Word Group Formation Methods

The relations between named entities are meant to make sense to news readers in the context of news material. Therefore, the word groups used to form relations, need to correspond to relevant *contexts* or specific stories in news material rather than to syntactic categories.

Brown clustering method assigns words to groups or classes according to their statistical behavior in a large body of text. Word groups or class labels are later used to learn a language model for the available body of text. Two words are put into one cluster if the words occurring in their vicinity are the same. For example, words ‘Thursday’ and ‘Friday’ will be put into one group because the same words occur around these two words in any large text corpora. We clustered words through Brown clustering and used these clusters as the group structure while learning the sparse logistic regression model for each entity. The resulting relations between named entities are of far lower quality than the relations discovered by other word clustering methods. Brown clustering groups words together based on their syntactic behavior in text corpus, not the news stories they describe.

Brown clustering would cluster words like ‘basketball’, ‘baseball’ and ‘football’ together, potentially merging different news stories about different sports events. The sample Brown cluster $\{shoot, dive, thrive, pass, advertise, crash, wear, propose, dismantle, adopt, amend, hunt\}$ from our dataset, seems to have grouped together verbs. The Brown cluster $\{jail, convict, gang, appeal, guilty, conviction, charge, sentence\}$ would result in relating entities from every law-and-order story to each other. On the other hand, the co-occurrence based cluster $\{media, outspoken, journalist, conviction\}$, for the same articles set, corresponds to one specific law-and-order story.

It is necessary for the word groups to correspond to specific news events or stories instead of syntactic categories, for extraction of named entities’ relation which are understandable in the context of news. Since the word groups formed by Brown clustering do not correspond to news stories, named entities’ relations based on those word groups do not make much sense in the context of news and do not fare well when evaluated through our evaluation criteria.

When we compared the performance of the three chosen word group formation methods, we observed that topic-based word groups tend to generate higher numbers of relations than co-occurrence- and keyword-based word groups, for the same value of γ . Only topic-based word groups are overlapping, thus forcing more named entities to share word groups with higher positive evidence. The mean *strength* assigned to the discovered relation is generally higher for topic-based word groups than all other group formation methods for comparable values of connectivity and quality measures (WLM and retrieval score). This is because threshold γ is set to a higher value for topic-based word groups to generate about the same number of relations as other methods.

4.4.3.5 Human Evaluation Study

We also conducted a human evaluation study of our system on the TIME dataset. The aim of this study is to compare the semantic network built by our system against that built by humans

when given the same set of news articles. In our human evaluation study, we fixed the duration of time slot to one day so as to limit the number of articles to a number easily readable by human judges. We selected two time slots, referred to as slot A and slot B, containing 10 and 8 articles, respectively. We presented the judges with a matrix of named entities for each time slot, such that each cell of the matrix corresponds to the pair of entities indicated by the row and the column. The judges were asked to read the articles for each time slot and mark in the matrix whether or not each pair of named entities is related based on the articles in the time slot. We collected observations from 16 judges. Likewise, NELasso was employed to automatically build semantic networks of named entities for slot A and slot B.

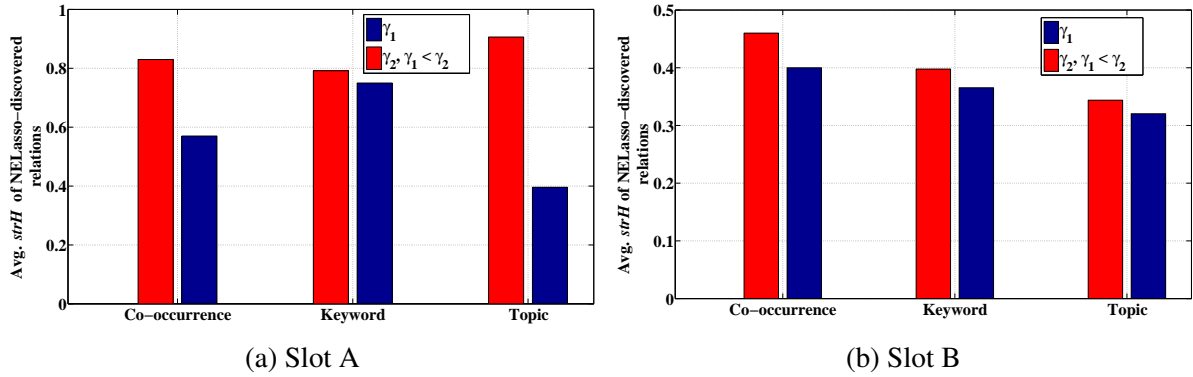


Figure 4.18: Human evaluation of NELasso; height of each bar represents the mean of human-assigned strength to the relations discovered by NELasso; Blue: $\gamma = \gamma_1$, Red: $\gamma = \gamma_2$ such that $\gamma_1 < \gamma_2$

The strength of a relation in human evaluation is estimated from the number of judges who mark that relation. When a relation is identified by many judges, it indicates that the relation is clear and strong enough to be recognized readily by humans. Accordingly, the strength, $strH_{ij}$, of a relation between entities e_i and e_j is defined as

$$strH(r_{ij}) = \frac{\text{No. of judges that identify } r_{ij}}{\text{Total no. of judges}} \quad (4.45)$$

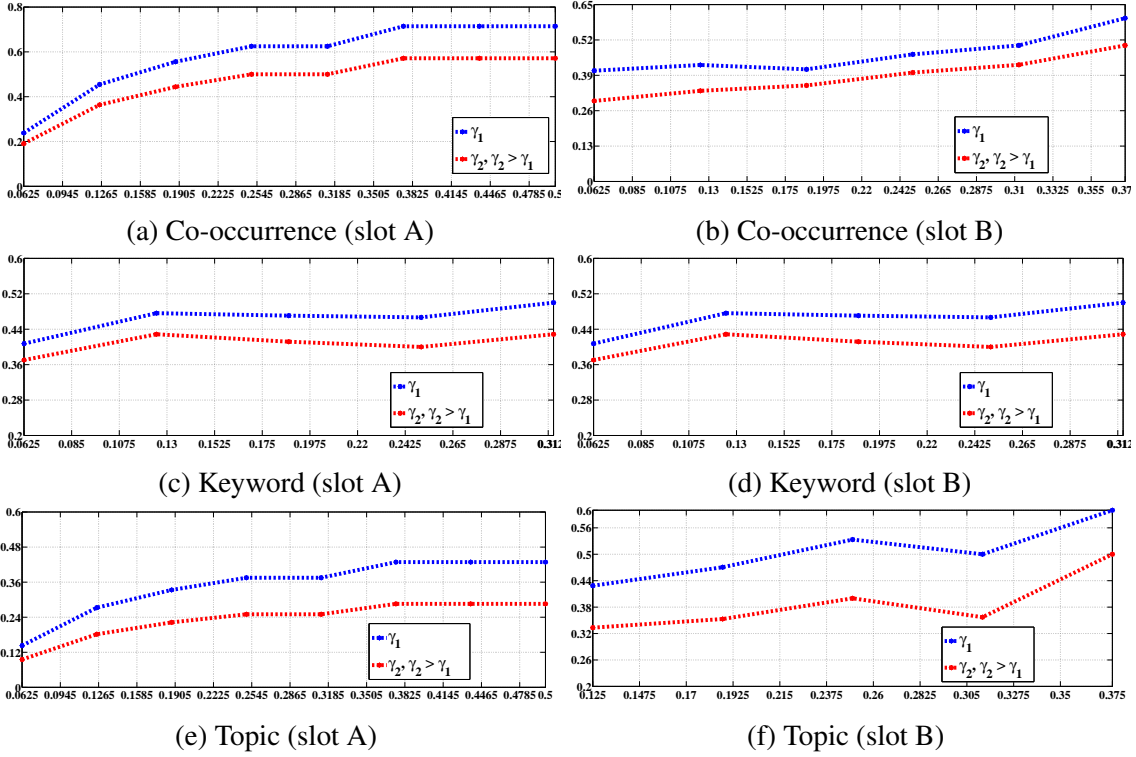
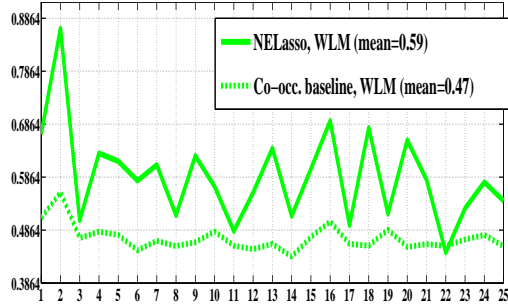


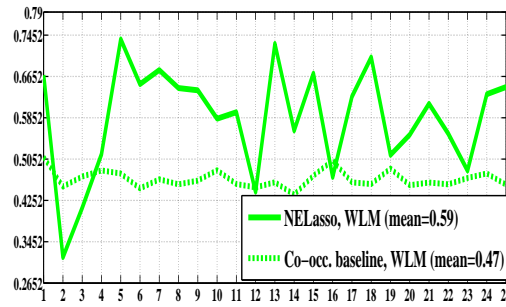
Figure 4.19: Fraction of human-identified relations discovered by NELasso; x-axis: Minimum $strH(r_{ij})$ of human-identified relations, y-axis: fraction of human-identified relations discovered by NELasso; Blue: $\gamma = \gamma_1$, Red: $\gamma = \gamma_2$ such that $\gamma_1 < \gamma_2$

Figure 4.18 shows that the $str(r_{ij})$ assigned to a relation between entities e_i and e_j by our system is a good indicator of $strH(r_{ij})$, i.e., the strength assigned to the relation by humans. As threshold γ is increased, NELasso identifies fewer but stronger relations. It is seen from Figure 4.18 that the fewer relations at higher γ also have higher mean $strH(r_{ij})$ than those selected by the lower γ value. NELasso-assigned relation *strength* correlates well with that assigned by the humans.

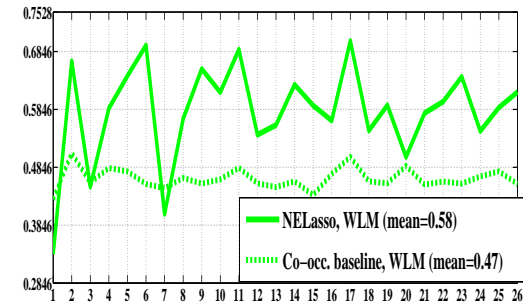
Our system is able to discover a higher fractions of human-identified relations with higher $strH(r_{ij})$ (Figure 4.19). The blue and red lines indicate the lower and the higher values of threshold γ , respectively. The horizontal axis shows the minimum $strH(r_{ij})$ of the relations identified by humans. In general, larger fractions of human-identified relations are discovered by NELasso (y-axis) with the increase in minimum $strH(r_{ij})$. This trend is more pronounced in slot A than in slot B.



(a) Baseline vs. Co-occurrence-based NELasso

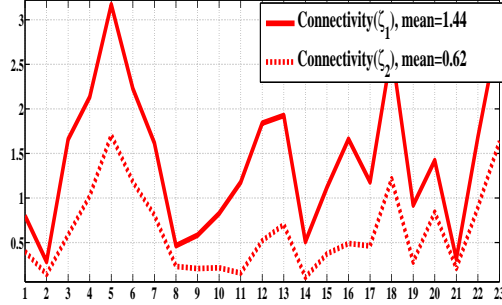


(b) Baseline vs. Keyword-based NELasso

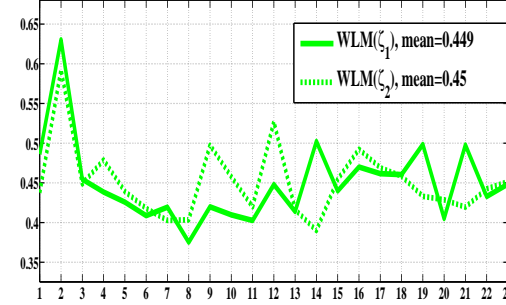


(c) Baseline vs. Topic-based NELasso

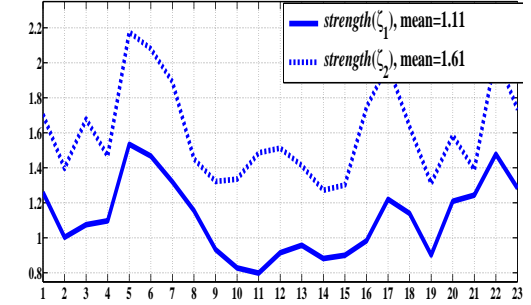
Figure 4.20: Comparison between co-occurrence-based baseline model and various configurations of NELasso using WLM (TIME dataset); x-axis: Time period (month), y-axis: mean WLM; Mean of each curve given in legends.



(a) Connectivity



(b) WLM



(c) Strength

Figure 4.21: Effect of threshold ζ of linear model baseline system on evaluation measures (TIME dataset); x-axis: Time period (month), y-axis: Evaluation measure (connectivity, mean WLM, mean retrieval score, mean *strength*); Solid line: $\zeta = \zeta_1$, Dotted line: $\zeta = \zeta_2$ where $\zeta_1 < \zeta_2$; Mean of each curve given in legends.

We also compute the Fleiss-kappa for human judges, which is an effective measure for inter-rater reliability[34]. Fleiss-kappa for slot A is 0.6 which reflects moderate-to-substantial inter-rater agreement. Fleiss-kappa for slot B is 0.37 which indicates fair agreement among the judges.

4.4.3.6 Co-occurrence-based Baseline Model

In this section, we present a baseline model for finding the relations between named entities. According to this model, a relation exists between two named entities when they co-occur in the same article. We compare the quality of relations found by NELasso and by this baseline model using WLM on different time periods of the TIME dataset. Figure 4.20 shows that the mean WLM for relations found in each month by any configuration of NELasso is much higher than that of the baseline model. This confirms that the straightforward method of constructing relations based on co-occurrence of entities in articles generates a large quantity of substandard relations with no information regarding their type or statistical signature.

4.4.3.7 Value of Sparse Group Learning

In this section, we address a fundamental question regarding our model: what is the benefit of sparse group learning over standard un-regularized learning?

To answer this question, we consider another baseline system that learns simple linear models for all named entities by using their positive and negative examples (i.e., articles) in equal numbers. This system learns a coefficient vector $\mathbf{p}_i \in \mathbb{R}^N$ for the i_{th} named entity, e_i . The relation $rel(e_i, e_j)$ between entities e_i and e_j is decided based upon the cosine similarity between \mathbf{p}_i and \mathbf{p}_j . If this similarity is greater than a threshold ζ , the system declares a relation r_{ij} between entities e_i and e_j with *strength* $str(r_{ij}) = \mathbf{p}_i^T \mathbf{p}_j$. There is no way of finding a meaningful *type* for this relation.

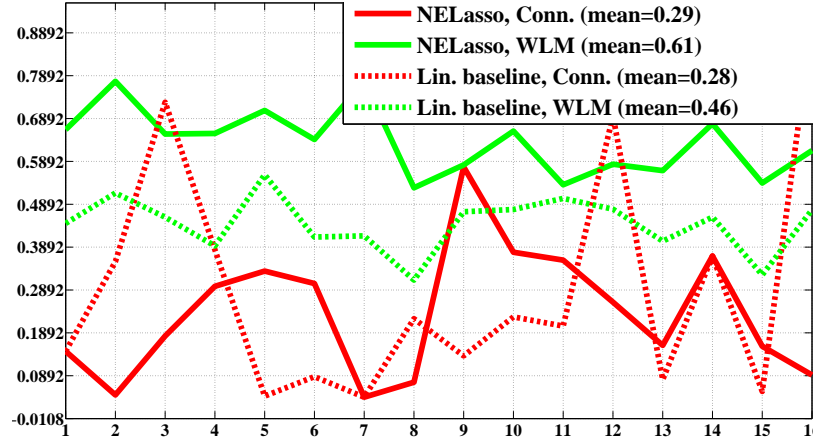


Figure 4.22: Comparison between NELasso and linear model baseline system (TIME dataset); x-axis: Time period (month), y-axis: Evaluation measure (connectivity, mean WLM); Mean of each curve given in legends.

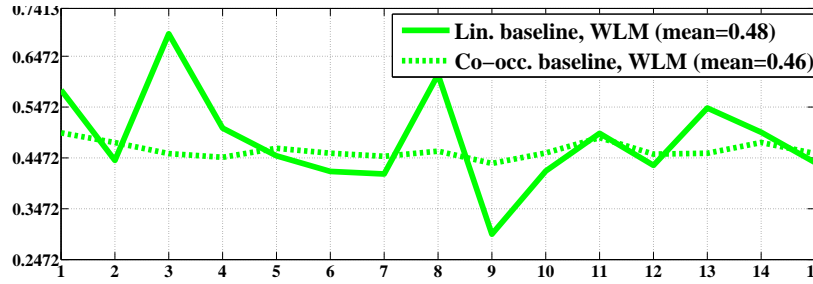


Figure 4.23: Comparison between co-occurrence and linear model based baselines on the basis of WLM (TIME dataset); x-axis: Time period (month), y-axis: mean WLM for relations; Mean of each curve in legend.

Figure 4.21 shows mean connectivity, WLM, and *strength* for semantic networks built by the linear model baseline system for each time interval of TIME dataset. Note that there is negligible change in mean WLM even after significant change in mean connectivity of the network. The difference between mean WLM as we change ζ is also negligible while the corresponding difference in NELasso is substantial (Figure 4.14). For time intervals where mean WLM changes with the increase in threshold ζ , often the change is negative indicating the deterioration in the quality of discovered relations. This implies that there is little correlation between the relation *strength* in the

linear model baseline system and the quality of the identified relations, as judged by WLM. This is the first advantage of employing group sparse learning in the our system.

Figure 4.22 highlights another advantage of NELasso over the simple linear model based baseline system. When the two systems find almost similar numbers of relations among named entities, the relations identified by sparse group learning are of much higher quality than those identified by the linear model based system. Furthermore, the sparse group learning based system assigns a meaningful relation *type* to each identified relation. No meaningful relation *type* can be identified in simple linear modeling based baseline.

The simple linear model performs only slightly better than the co-occurrence-based model presented in Section 4.4.3.6. Figure 4.23 shows that the mean WLM of the relations found by the linear model is higher than that of co-occurrence-based model for a few time intervals only. In comparison, NELasso performs consistently better than both baseline models in terms of the mean WLM as shown in Figure 4.20.

4.4.3.8 Sensitivity Analysis

Our system requires tuning of a few parameters before its execution. The parameters include the weights λ_1 and λ_2 assigned to the two penalty terms involved in the group sparse logistic regression model and the threshold γ on the relation *strength* for its selection. We experimentally study the effects of these parameters on the output of the system. Furthermore, we also study the impact of sampling of negative articles on system output.

The parameter γ controls the selection of relations such that only relations with $str_k(r_{ij}) = t_k^{e_i} \times t_k^{e_j}$ where $t_k^{e_i} > \gamma$ and $t_k^{e_j} > \gamma$ are included in the output (refer to Definition 1 for details). As γ is increased, fewer relations of higher *strength* (quality) are selected for inclusion in the semantic

network. Moreover, the relations with high *strength* are unaffected by even a significant increase in γ . This trend is discussed in detail in Section 4.4.3.3 (see Figures 4.14 and 4.16).

We observe that the increase in values of λ_1 and λ_2 has the same effect as increase in the value of γ . As λ_1 and λ_2 are increased, more emphasis is put on sparsity of the logistic regression model, i.e., the entries of coefficient vector \mathbf{x} become smaller and a larger number of them are set to 0. Since the relations between named entities are decided based on sum of entries x_n of \mathbf{x} such that n_{th} word belongs to a certain word group (see Section 4.4.1.2), fewer relations are discovered with the increase in values of these parameters. But, the relations with high *strength* are unaffected as the sum of coefficients of the named entities will be higher than those for other entities. Of course, the threshold γ has to be adjusted downward since the absolute strength value will be lower.

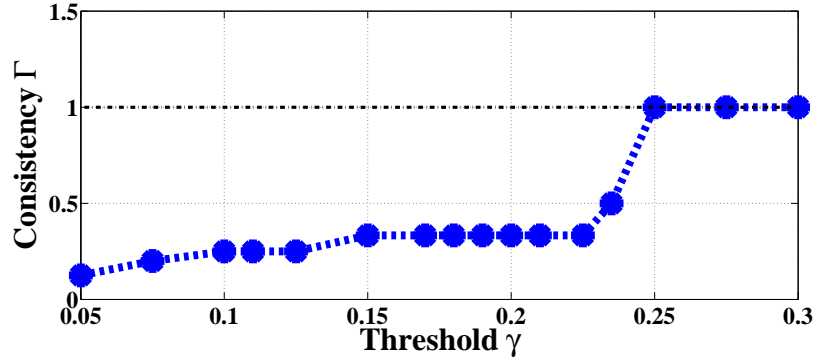


Figure 4.24: Effects of threshold γ on the consistency of system's output (TIME dataset)

While learning the group sparse logistic regression model for a named entity, our system randomly selects a set of articles that do not mention the named entity (since the number of articles that mention an entity is much smaller than those that do not). We study the sensitivity of the system's output to selection of different set of negative examples by evaluating the output from five runs of the system. Each time, the system randomly selects sets of negative examples for each entity. Let \mathcal{R}_t be the set of relations discovered in the t_{th} iteration out of a total of T iterations, then the

consistency of the systems output is defined as the ratio of the number of common relations found in all iteration to the number of unique relations found in all iterations. Its maximum value is 1, i.e., all relations are discovered in all iterations.

$$consistency = \frac{|\cap_{t=1}^T \mathcal{R}_t|}{|\cup_{t=1}^T \mathcal{R}_t|} \quad (4.46)$$

Figure 4.24 shows the effect of γ on system consistency. It is observed that as γ is increased the systems output becomes more and more consistent until its consistency reaches 1. This implies that relations of high *strength* are consistently discovered in all iterations despite variations in selection of negative examples.

4.4.3.9 Time Complexity and Scalability

NELasso is not only effective but also time efficient and scalable to large-scale applications of identifying relations among named entity from published news articles automatically. The system identifies word groups once for a set of articles and uses them while learning a sparse logistic regression model for each named entity mentioned in that set of articles. Our system takes on average 0.05 seconds to process one named entity on a machine with 3.40 GHz processor and 32 GB memory. It is clear that our system can scale up easily to practical settings involving large sets of news articles collected from multiple sources on daily basis.

4.5 Conclusion

In this chapter, we explained our ideas about automatic understanding of valuable semantic relations between visual and textual data of news collections. News dataset is a classic example of multi-modality datasets, containing long and short text, images, structured data as well as times-

tamps. Such a dataset provides an excellent opportunity for expansion and implementation of the core idea of our dissertation, i.e., cross-media semantic relation building with a focus on image-text relations. Such datasets are rich in information content. Therefore, important real world applications like search and retrieval engines, information summarization and visualization tools, and query-answering frameworks, need to be able to automatically understand semantic information encoded in associations between various data types involved in such datasets. For thorough experimentation, we collected a large dataset of news image, along with their captions, news articles, and metadata. This dataset has the potential to become the benchmark for evaluation of various systems involving multi-modality datasets such as search engines, query-answering tools, event tracking and summarization frameworks.

We devised an automatic image annotation system for news images that aims at matching real world ground truth descriptions of these images. Since real world image captions involve hints to the *context* of images, the proposed system collects and incorporates semantic contextual cues from all available sources of different data modalities, i.e., semantic scene category of images, topics discussed in news articles, news category labels and keywords. Semantic information is propagated between these heterogeneous data sources by employing probability space as common representation space. We also devised a framework to generate sentence-like captions for news images that employs the annotations predicted for such images and an *extractive* framework to extract the best caption from the associated news article. *Extractive* framework is also sensitive to the contextual cues associated with news items.

To fully utilize the potential of long sequences of text available with news images, i.e., news articles, we devised a system to understand semantic relations between named entities. Named entities are the most important linguistic features in terms of search and retrieval. Such entities are also linked to images directly through identification of people, logos and landmarks. Such image-text relations can be semantically enriched if semantic relations between named entities are

presented in a machine understandable way. Our system is focused on estimating semantic context of named entities through sparse structured modeling of their occurrence in news articles using vocabulary words as predictors. Semantic topics are automatically defined from the group structure found in the set of vocabulary words. Common semantic context of two named entities in any given time period, indicates a semantic relation between the two entities. Our semantic topic extraction approach and sparse structured modeling scheme enables our system to assign *type* and *strength* to each relation. We evaluated these relations based on their verification from independent sources, as well as their value to retrieval and recommendation tools. In comparison to other systems devised to understand relations between named entities, our system is unsupervised, does not need seed relational tuples, hand-crafted rules or external databases, and can extract unlimited number of relation *types*.

The systems that we presented in this chapter has vast potential to be incorporated in any real world search, retrieval, summarization or story tracking tool dealing with databases containing multiple modalities of data. Vast scope of application is one of our major concerns while designing any systems. Therefore, our systems do not require manually crafted input or supervision. Since semantic topic extraction is an important part of our work, we devise frameworks to automatically extract such topics from available training data instead of manual identification of such topics. We devise various frameworks to establish semantic relations for data items, to suit the needs of any given setup. Our frameworks include techniques from the fields of text mining, language processing and image processing.

CHAPTER 5: CONCLUSION

In this chapter, we summarize the motivation behind the ideas that we discussed in previous chapters, as well as presenting the conclusions drawn from our extensive experimental evaluation process.

We are getting access to larger and more diverse datasets than ever before, through the internet. It is necessary to build systems that can automatically understand semantic information hidden inside such datasets. Such intelligent systems can help humans take advantage of the available datasets through aiding in search, information retrieval and database organization tasks.

Traditionally, search and retrieval systems were focused on a single data modality, e.g., measuring similarity between textual query and documents in any textual database. Heterogeneous datasets involving multiple data modalities such as images, text, audio, video, discrete domain metadata, etc., are now available through sources like social media websites and online news media outlets. To perform meaningful search and organizational operations on such datasets, intelligent systems need to automatically understand semantic relation across different data modalities.

Images and text are two very common and important data modalities. Automatic understanding of image-text relations is a challenging task but such understanding can be of tremendous help to image search engines, query-answering tools, information summarization and visualization systems involving databases containing visual and textual data. Prediction of suitable word annotations for images is called automatic image annotation. *Semantic gap*, i.e., the lack of the correlation between visual and textual features is the main challenge to machine understanding of image-text relations.

This dissertation deals with the challenge of machine understanding of cross-modality datasets,

with a focus on image-text relations. Our hypothesis is that it is crucial to understand the semantic context of the information present in the available database to build meaningful cross-modality relations. There may be no strict correspondence between low-level visual features and words in general, but meaningful relations between images and their given semantic *contexts* can be built. Such semantic contextual relations contain valuable information that can be employed as prior knowledge while developing image-word relations. For example, if an image is known to show characteristics of an ‘open country’-type semantic scene, words like ‘green’, ‘grass’, ‘field’ are highly likely to be associated with this image (Section 3.1 of Chapter 3). Our work is focused on understanding semantic contextual relations of images without requiring any external resource or additional input, or imposing any special restriction on the available dataset.

We presented three automatic image annotation systems in Chapter 3. These systems employ only the available training data to extract semantic *context* types in terms of semantic scene categories and word-groups corresponding to visual themes, and to quantify the relations between test images and semantic *context* types. We explored different models to incorporate such semantic contextual relations in the process of automatic image annotation. The performance of these three systems when compared against that of a wide range of previously proposed system, speaks to the validity of the main hypothesis of this dissertation, i.e., meaningful cross-modality relations can be automatically developed if semantic background of the available information is quantified.

After verifying the validity of our hypothesis, we turned our attention to an even more diverse dataset. We collected a large dataset of news images, along with their caption, corresponding news articles, article keywords and titles, news category labels as well as timestamps. We closely studied the nature of ground truth captions of news images that we downloaded in comparison to the nature of ground truth captions available in popular image annotation datasets, like IAPR TC-12, Flickr30K and MSCOCO. We observed that previously available image annotation datasets contain ‘artificial’ image captions, in the sense that these captions were not used to describe these

images in the real world. For example, MSCOCO and Flickr30K contain images downloaded from the Flickr website. Instead of using their actual captions as the ground truth, human annotators with no knowledge regarding the background of these images were asked to write their captions. Hence, these artificially-produced captions describe visual contents of images in basic terms such as names of objects and actions presented in images. On the other hand, real world captions available in our dataset almost always refer to the *context* of images or the stories behind these images. Correlation between visual and textual features seems to be even weaker, and hence the *semantic gap* even wider for this dataset. We described the details of our dataset as well as our observations regarding the nature of the real world image captions in Chapter 4.

Since the real world news image captions may include hints to the information outside of the visual contents of the images, an annotation prediction system also needs to understand the *context* of these images with the help of every available source of information both intrinsic and extrinsic to the images. We identified four different sources of such contextual information, i.e., the semantic scene properties of images, contents of the articles associated with images, news category labels and the keywords assigned to news items. These sources belong to different data modalities. We chose the probability space as the common representation space for contextual information collected from these heterogeneous sources. We devised methods to estimate association between news images and the *context* categories of types defined by all of these sources. We also devised a generative model conditioned over the semantic contextual information of the test news image to estimate the joint probability between the test image and the vocabulary words. Top words with respect to this joint probability are chosen as the annotations for the news images.

We also devised an *extractive* framework for generation of sentence-like image captions. Our system estimates the word distribution for the ideal caption of the image in light of the contents of the image and its associated article in a *context*-sensitive manner. Since a news article is available with the news image, our framework employs the estimated word-distribution to select the

best sentence from the pool of sentences of the article to be used as image caption. Since news article contain grammatically correct text, our technique is guaranteed to produce grammatically correct sentences as opposed to some *abstractive* techniques that may produce incomprehensible sentences.

We thoroughly evaluated our image annotations and caption generation frameworks against a wide variety of simple baselines and previously proposed methods. Most of the previously proposed annotation and caption generation systems have no provision to incorporate semantic information collected from auxiliary information sources like news articles and metadata. Our annotation and caption generations models outperform such methods. Our models even outperform the few previously proposed models which incorporate information from news articles. Deep neural network based caption generation methods rely on ImageNet-trained convolutional neural networks (CNN) to produce image representations. Such systems have been very successful in automatic description generation for standard image annotation datasets containing artificial ground truth. Our experiments show that such CNN-produced image representations have little correlation with ground truth descriptions of news images. We discussed our annotation and caption generation models and the analysis of their performance in Sections 4.2 and 4.3 of Chapter 4.

Vast amount of semantic information is hidden in free-flowing long sequences of news articles. The relation between news images and this semantic information is somewhat indirect but still very important. To build concrete image-text relations out of such semantic information, information encoded in articles need to be distilled to a form where it is directly identifiable in references to images. Names of people, places and organizations, i.e., the named entities, constitute the most-often used query terms for news and blogs datasets. Hence, these named entities are an information-rich linguistic feature of news articles. Facial recognition, logo detection and landmark identification systems have been developed to link these linguistic features of the text and images with each other. Such image-text relations can be further enriched if the system can also

access information regarding these images from the news articles. We developed a system to automatically extract semantically meaningful relations between named entities from the articles. We presented our model in Section 4.4 of Chapter 4.

We employed the main hypothesis of this dissertation for named entities semantic relations extraction as well. We explored multiple methods to build word-groups such that each group defines some type of semantic *context*. We modeled named entities' occurrence in articles through sparse logistic regression with words of articles as predictors and word-groups as inherent structure among predictors. Such modeling identifies words and word-groups that correlate with the occurrence of each named entity. We used entity-word group relations to *quantify* relations between named entities. Common word-groups between two named entities define the *type* of the relationship. Such relations do not only enrich image-entity relations determined by facial recognition, logo detection and landmark identification systems, but can also aid news recommendation or retrieval type systems. The results of our evaluation experiments prove that named entities relations extracted by our systems are not only valid but are also effective for news retrieval or recommendation tasks.

Cross-modality relation extraction systems presented in this dissertation have vast potential for application in the fields of data-driven advertisement and multi-modality data analytics. Online targeted advertisement can be made much more effective if the system can automatically understand the semantic meaning and implications of information available on users' social media account in the form of various data modalities. It is hard for humans to scan information from many news media outlets simultaneously, in a timely manner. A system will be immensely useful for both the news readers and the news editors if it can automatically build a summary of information or a timeline for every news event, find similar events reported in the past, or answer questions about people or places involved in any event. Such a system must develop semantic understanding of multi-modality news media contents. All of these systems present potential application scenarios

for the systems we developed for machine understanding of cross-media semantic relations such as image-text relations.

The main hypothesis of this dissertation, i.e., the semantic contextual information is crucial to the automatic understanding of cross-modality relations, can be generalized to include various data modalities. This quality was demonstrated when we included four different types of data to define semantic *context* to define image-text relations for news datasets. In future, we would like to include data modalities like video and audio that require time-sensitive processing. We dealt with the time component of news articles for named entities relation extraction system in a rather crude way. Sequential or time-dependent modeling of semantic contextual relations need to be developed. Such modeling can be immensely beneficial for processing vast amount of time-sensitive data available in the form of surveillance and social media videos and audio as well as social-media run-time responses to events like sports competitions.

We analyzed the performance of ours and various other image description generation system such as deep convolutional neural networks. We observed that the performance of any caption generation system needs to be trained over the a dataset which is characteristically similar to the testing data. This is the reason the state-of-the-art deep neural networks fail to generate reasonable captions for news images. Such deep networks rely on ImageNet database to learn image representations. Ground truth captions of news images are characteristically very different from simplistic labels of ImageNet database. This phenomenon is usually studied in the fields of *transfer learning* or *domain adaptation*. In future, we would like to study *domain adaptation* and *transfer learning* for the problem of automatic image caption generation. If the information from a characteristically different *source* or the training domain can be adapted to fit the *target* or the testing domain, it can open vast possibilities for effectively training of caption generation frameworks as well as other cross-media relations extraction systems.

LIST OF REFERENCES

- [1] Rabeeh Abbasi, Sergey Chernov, Wolfgang Nejdl, Raluca Paiu, and Steffen Staab. Exploiting flickr tags and groups for finding landmark photos. In *Advances in Information Retrieval*, pages 654–661. Springer, 2009.
- [2] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*.
- [3] Brett W Bader, Michael W Berry, and Murray Browne. Discussion tracking in enron email using parafac. In *Survey of Text Mining II*, pages 147–163. Springer, 2008.
- [4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- [5] Tamara L Berg, Alexander C Berg, Jaety Edwards, Michael Maire, Ryan White, Yee-Whye Teh, Erik Learned-Miller, and David A Forsyth. Names and faces in the news. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–848. IEEE, 2004.
- [6] Irving Biederman. Aspects and extensions of a theory of human image understanding. *Computational processes in human vision: An interdisciplinary perspective*, pages 370–428, 1988.
- [7] Christopher M Bishop. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

- [8] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, 2003.
- [9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [10] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022, 2003.
- [11] Danushka Tarupathi Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Relational duality: Unsupervised extraction of semantic relations between entities on the web. In *Proceedings of the 19th international conference on World wide web*.
- [12] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.
- [13] Liujuan Cao, Rongrong Ji, Yue Gao, Yi Yang, and Qi Tian. Weakly supervised sparse coding with geometric consistency pooling. In *CVPR, 2012*.
- [14] Min-Ta Chang and Shu-Yuan Chen. Deformed trademark retrieval based on 2d pseudo-hidden markov model. *Pattern Recognition*, 34(5):953–967, 2001.
- [15] BILIAN CHEN, ZHENING LI, and SHUZHONG ZHANG. On tensor tucker decomposition: The case for an adjustable core size.
- [16] David M Chen, Georges Baatz, Kevin Köser, Sam S Tsai, Ramakrishna Vedantham, Timo Pylvä, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, et al. City-scale landmark identification on mobile devices. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 737–744. IEEE, 2011.

- [17] David M Chen, Georges Baatz, Kevin Köser, Sam S Tsai, Ramakrishna Vedantham, Timo Pylvä, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, et al. City-scale landmark identification on mobile devices. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 737–744. IEEE, 2011.
- [18] Minmin Chen, Alice Zheng, and Kilian Weinberger. Fast image tagging. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1274–1282, 2013.
- [19] Xiangyu Chen, Xiaotong Yuan, Shuicheng Yan, Jinhui Tang, Yong Rui, and Tat-Seng Chua. Towards multi-semantic image annotation with graph regularized exclusive group lasso. In *Proceedings of the 19th ACM international conference on Multimedia*.
- [20] Dongjin Choi and Pankoo Kim. Automatic image annotation using semantic text analysis. In *Multidisciplinary Research and Practice for Information Systems*, pages 479–487. Springer, 2012.
- [21] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.
- [22] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. On the best rank-1 and rank-(r_1, r_2, \dots, r_n) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000.
- [23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- [24] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, 2004.

- [25] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014.
- [26] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, 2008.
- [27] Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, pages 100–110. ACM, 2004.
- [28] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, 2011.
- [29] Wei Fan, Jun Sun, Satoshi Naoi, Akihiro Minagawa, and Yoshinobu Hotta. Natural scene logo recognition by joint boosting feature selection in salient regions, 2011.
- [30] SL Feng, R Manmatha, and V Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition*, pages 1002–1009, 2004.
- [31] Yansong Feng and Mirella Lapata. Automatic image annotation using auxiliary text information. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 272–280, 2008.
- [32] Yansong Feng and Mirella Lapata. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 831–839, 2010.

- [33] Yansong Feng and Mirella Lapata. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812, 2013.
- [34] L Fleiss, Bruce Levin, and Myunghee Cho Paik. The measurement of interrater agreement. In *Statistical methods for rates and proportions (2nd ed)*, 1981.
- [35] Hao Fu, Qian Zhang, and Guoping Qiu. Random forest for image annotation. In *Computer Vision–ECCV 2012*, pages 86–99. 2012.
- [36] Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Multi-layer group sparse coding for concurrent image classification and annotation. In *CVPR, 2011*.
- [37] Shenghua Gao, Ivor Wai-Hung Tsang, Liang-Tien Chia, and Peilin Zhao. Local features are not lonely laplacian sparse coding for image classification. In *CVPR, 2010*.
- [38] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *Computer Vision–ECCV 2014*, pages 529–545. Springer, 2014.
- [39] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *IEEE 12th International Conference on Computer Vision*, pages 309–316, 2009.
- [40] Yahong Han, Fei Wu, Qi Tian, and Yueting Zhuang. Image annotation by input-output structural grouping sparsity. *IEEE Transactions on Image Processing*.
- [41] Zhifeng Hao, Lifang He, Bingqian Chen, and Xiaowei Yang. A linear support higher-order tensor machine for classification. *IEEE Transactions on Image Processing*, 22(7):2911–2920, 2013.

- [42] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 415. Association for Computational Linguistics, 2004.
- [43] Toru Hirano, Yoshihiro Matsuo, and Genichiro Kikui. Detecting semantic relations between named entities in text using contextual features. In *Proceedings of the 45th annual meeting of the ACL on Interactive poster and demonstration sessions*.
- [44] Keiichiro Hoashi, Toshiaki Uemukai, Kazunori Matsumoto, and Yasuhiro Takishima. Constructing a landmark identification system for geo-tagged photographs based on web data analysis. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 606–609. IEEE, 2009.
- [45] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pages 853–899, 2013.
- [46] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.
- [47] Yongzhen Huang, Zifeng Wu, Liang Wang, and Tieniu Tan. Feature coding in image classification: A comprehensive study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):493–506, 2014.
- [48] Mariya Ishteva, Lieven De Lathauwer, P-A Absil, and Sabine Van Huffel. Differential-geometric newton method for the best rank-(r_1, r_2, r_3) approximation of tensors. *Numerical Algorithms*, 51(2):179–194, 2009.

- [49] Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126, 2003.
- [50] Yuzhe Jin, Emre Kiciman, Kuansan Wang, and Ricky Loynd. Entity linking at the tail: sparse signals, unknown entities, and phrase models. In *Proceedings of the 7th ACM international conference on Web search and data mining*.
- [51] Yuzhe Jin, Kuansan Wang, and Emre Kiciman. Sparse lexical representation for semantic entity resolution. In *IEEE international conference on Acoustics, speech and signal processing (ICASSP), 2013*.
- [52] K Sparck Jones et al. Automatic summarizing: factors and directions. *Advances in automatic text summarization*, pages 1–12, 1999.
- [53] Khurum Nazir Junejo and Asim Karim. A robust discriminative term weighting based linear discriminant method for text classification. In *IEEE International Conference on Data Mining (ICDM)*, pages 323–332, 2008.
- [54] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*, 2014.
- [55] Anastasios Kesidis and Dimosthenis Karatzas. Logo and trademark recognition. In *Handbook of Document Image Processing and Recognition*, pages 591–646. Springer, 2014.
- [56] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.
- [57] Ryan Kiros and Csaba Szepesvári. Deep representations and codes for image auto-annotation. In *Advances in Neural Information Processing Systems*, pages 908–916, 2012.

- [58] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.
- [59] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [60] Gerald Kuhne, Joachim Weickert, Oliver Schuster, and Stephan Richter. A tensor-driven active contour model for moving object segmentation. In *Proceedings of IEEE International Conference on Image Processing*, 2001.
- [61] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating simple image descriptions. In *IEEE conference on Computer vision and pattern recognition (CVPR)*, pages 1601–1608, 2011.
- [62] Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 359–368. Association for Computational Linguistics, 2012.
- [63] Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. Generalizing image captions for image-text parallel corpus. In *ACL (2)*, pages 790–796, 2013.
- [64] Victor Lavrenko, R Manmatha, and J Jeon. A model for learning the semantics of pictures. In *Proceedings of the 17th International Conference on Neural Information Processings Systems (NIPS)*, 2003.

- [65] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [66] Chee Wee Leong and Rada Mihalcea. Explorations in automatic image annotation using textual features. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 56–59, 2009.
- [67] Chee Wee Leong, Rada Mihalcea, and Samer Hassan. Text mining for automatic image tagging. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 647–655, 2010.
- [68] Xirong Li, Cees GM Snoek, and Marcel Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322, 2009.
- [69] Zhixin Li, Zhongzhi Shi, Weizhong Zhao, Zhiqing Li, and Zhenjun Tang. Learning semantic concepts from image database with hybrid generative/discriminative approach. *Engineering Applications of Artificial Intelligence*, 26(9), 2013.
- [70] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. 2014.
- [71] Xiaobai Liu, Bin Cheng, Shuicheng Yan, Jinhui Tang, Tat Seng Chua, and Hai Jin. Label to region by bi-layer sparsity priors. In *Proceedings of the 17th ACM international conference on Multimedia*.
- [72] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- [73] Haiping Lu, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos. A survey of multilinear subspace learning for tensor data. *Pattern Recognition*, 44(7):1540–1551, 2011.

- [74] Z. Lu, H. H. S. Ip, and Y. Peng. Contextual kernel and spectral methods for learning the semantics of images. *IEEE Transactions on Image Processing*, 20(6):1739–1750, 2011.
- [75] Zhiwu Lu, Peng Han, Liwei Wang, and Ji-Rong Wen. Semantic sparse recoding of visual content for image applications. *Image Processing, IEEE Transactions on*, 24(1):176–188, 2015.
- [76] Zhigang Ma, Feiping Nie, Yi Yang, Jasper RR Uijlings, and Nicu Sebe. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Transactions on Multimedia*, 14(4):1021–1030, 2012.
- [77] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. A new baseline for image annotation. In *Proceedings of the 10th European conference on Computer Vision: Part III*, pages 316–329. 2008.
- [78] Inderjeet Mani and Mark T Maybury. *Advances in automatic text summarization*, volume 293. MIT Press, 1999.
- [79] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan Yuille. Deep captioning with multi-modal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.
- [80] Rebecca Mason and Eugene Charniak. Apples to oranges: Evaluating image annotations from natural language processing systems. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 172–181, 2012.
- [81] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on Natural language processing of the AFNLP*.

- [82] Gilad Mishne and Maarten De Rijke. A study of blog search. In *Advances in information retrieval*, pages 289–301, 2006.
- [83] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, 2012.
- [84] Sean Moran and Victor Lavrenko. Optimal tag sets for automatic image annotation. In *Proceedings of the British Machine Vision Conference*, 2011.
- [85] Aude Oliva and Philippe G Schyns. Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive psychology*, 34(1):72–107, 1997.
- [86] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
- [87] Joseph Olive. Global autonomous language exploitation (gale). *DARPA/IPTO Proposer Information Pamphlet*, 2005.
- [88] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, 2011.
- [89] Derya Ozkan and Pınar Duygulu. A graph based approach for naming faces in news photos. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1477–1482. IEEE, 2006.

- [90] Kannappan Palaniappan, Ilker Ersoy, Guna Seetharaman, Shelby R Davis, Praveen Kumar, Raghuveer M Rao, and Richard Linderman. Parallel flux tensor analysis for efficient moving object detection. In *Proceedings of the 14th IEEE International Conference on Information Fusion (FUSION)*, pages 1–8, 2011.
- [91] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.
- [92] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
- [93] Anh-Huy Phan, Andrzej Cichocki, and Petr Tichavsky. On fast algorithms for orthogonal tucker decomposition. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2014.
- [94] Mary C Potter. Meaning in visual search. *Science*, 1975.
- [95] Duangmanee Putthividhy, Hagai Thomas Attias, and Srikantan S Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3408–3415, 2010.
- [96] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *IEEE 11th international conference on Computer vision (ICCV)*, pages 1–8, 2007.
- [97] Stephen Roller and Sabine Schulte Im Walde. A multimodal lda model integrating textual, cognitive and visual modalities. *Journal of Artificial Intelligence Research*, 18:1–44, 2003.

- [98] Benjamin Rosenfeld and Ronen Feldman. Clustering for unsupervised relation identification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 411–418, 2007.
- [99] Michael Rubinstein, Ce Liu, and William T Freeman. Annotation propagation in large image databases via dense image correspondence. In *Proceedings of the 12th European conference on Computer Vision-Volume Part III*, pages 85–99. 2012.
- [100] M. Rusiol and J. Llads. Logo spotting by a bag-of-words approach for document categorization. In *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, pages 111–115, July 2009.
- [101] Hichem Sahbi, Lamberto Ballan, Giuseppe Serra, and Alberto Del Bimbo. Context-dependent logo matching and retrieval. Technical report, TELECOM ParisTech, Tech. Rep, 2010.
- [102] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, 2012.
- [103] Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. Linden: linking named entities with knowledge base via semantic knowledge. In *Proceedings of the 21st international conference on World Wide Web*.
- [104] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [105] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

- [106] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [107] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.
- [108] Chuan Sun, Marshall Tappen, and Hassan Foroosh. Feature-independent action spotting without human localization, segmentation, or frame-wise tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
- [109] Sean Szumlanski and Fernando Gomez. Automatically acquiring a semantic network of related concepts. In *Proceedings of the 19th ACM international conference on Information and knowledge management*.
- [110] Xiaoyang Tan, Songcan Chen, Zhi-Hua Zhou, and Fuyan Zhang. Face recognition from a single image per person: A survey. *Pattern recognition*, 39(9):1725–1745, 2006.
- [111] Amara Tariq and Hassan Foroosh. Scene-based automatic image annotation. In *Proceedings of the 2014 IEEE international conference on Image processing*, 2014.
- [112] Amara Tariq and Hassan Foroosh. Feature-independent context estimation for automatic image annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [113] Amara Tariq and Hassan Foroosh. T-clustering: Image clustering by tensor decomposition. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 4803–4807. IEEE, 2015.
- [114] Amara Tariq and Hassan Foroosh. A context-driven extractive framework for generating realistic image descriptions. In *IEEE Transaction on Image Processing*, 2016.

- [115] Amara Tariq and Asim Karim. Fast supervised feature extraction by term discrimination in information pooling. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2233–2236, 2011.
- [116] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*.
- [117] Peter D Turney, Patrick Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*.
- [118] Yoshitaka Ushiku, Tatsuya Harada, and Yasuo Kuniyoshi. Efficient image annotation for automatic sentence generation. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 549–558. ACM, 2012.
- [119] M Alex O Vasilescu and Demetri Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *Computer Vision–ECCV 2002*.
- [120] Yashaswi Verma and CV Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *Proceedings of the 12th European conference on Computer Vision–Volume Part III*, pages 836–849. 2012.
- [121] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.
- [122] I Witten and David Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI workshop on Wikipedia and artificial intelligence: an Evolving synergy*, pages 25–30, 2008.
- [123] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE PAMI*, 2009.

- [124] Fei Wu and Daniel S Weld. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, 2010.
- [125] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.
- [126] Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 444–454, 2011.
- [127] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.
- [128] Qibin Zhao, Guoxu Zhou, Liqing Zhang, and Andrzej Cichocki. Tensor-variate gaussian processes regression and its application to video surveillance. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
- [129] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.
- [130] Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. Statsnowball: a statistical approach to extracting entity relationships. In *Proceedings of the 18th international conference on World wide web*.